Hypothesis Testing in GWAS and Statistical Issues with Compensation in Clinical Trials

Hypothesis Testing in GWAS and Statistical Issues

with Compensation in Clinical Trials

A dissertation presented

by

David Michael Swanson

to

The Department of Biostatistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biostatistics

Harvard University

Cambridge, Massachusetts

April 2013

Dissertation Advisors: Rebecca Betensky and Christoph Lange David Michael Swanson

Hypothesis Testing in GWAS and Statistical Issues with Compensation in Clinical Trials

Abstract

We first show theoretically and in simulation how power varies as a function of SNP correlation structure with currently-implemented gene-based testing methods. We propose alternative testing methods whose power does not vary with the correlation structure. We then propose hypothesis tests for detecting prevalence-incidence bias in case-control studies, a bias perhaps overrepresented in GWAS due to currently used study designs. Lastly, we hypothesize how different incentive structures used to keep clinical trial participants in studies may interact with a background of dependent censoring and result in variation in the bias of the Kaplan-Meier survival curve estimator.

Table of Contents

Properties of permutation tests and alternative methods for gene-based testing

David M. Swanson*, Deborah Blacker, Taofik Al-Chawa,

Kerstin U. Ludwig, Elisabeth Mangold, Christoph Lange

Abstract

The advent of genome-wide association studies has led to many novel disease-SNP associations, opening the door to focused study on their biological underpinnings. Because of the importance of analyzing these associations, numerous statistical methods have been devoted to them. However, fewer methods have attempted to associate entire genes or genomic regions with outcomes, which is potentially more useful knowledge from a biological perspective and those methods currently implemented are often permutation-based. One property of some permutation-based tests is that their power varies as a function of whether significant markers are in regions of linkage disequilibrium (LD) or not, which we show from a theoretical perspective. We therefore develop two methods for quantifying the degree of association between a genomic region and outcome, both of whose power does not vary as a function of LD structure. One method uses dimension reduction to "filter" redundant information when significant LD exists in the region, while the other controls for LD by scaling marker Z-statistics using knowledge of the correlation matrix of markers. One advantage of the former test is power gains due to dimension reduction, while an advantage of the latter test is that the full data set is not needed, but just summary Z-statistics from univariate regressions of markers and the underlying correlation structure of those markers which are publicly available in some cases. We show how to modify the latter test when the correlation structure of markers is imperfectly known in order to protect the type 1 error rate. We apply these methods to sequence data of oral cleft and compare our results to previously proposed gene tests, in particular permutation-based ones. We find a significant association in the sequence data between the 8q24 region and oral cleft using our dimension reduction approach and a borderline significant association using the summary-statistic based approach. We also implement the summary-statistic test using Z-statistics from an already-published GWAS of Chronic Obstructive Pulmonary Disorder (COPD) and correlation structure obtained from HapMap. We experiment with the modification of this test because the correlation structure is assumed imperfectly known.Dimension reduction; Eigenvector; Gene-based testing; Permutation tests.

1 Introduction

The focus in genetic association studies has been on uncovering loci that are risk factors for an outcome, be it binary or continuous, or markers in linkage disequilibrium (LD) with those causal loci. Increasingly, however, gene-based tests are coming to the forefront, especially as sequencing technologies mature and grow cheaper [8, 3, 14]. Gene-based tests are useful to provide insight into whether a region of the genome has a significant association with some outcome and for inter-gene significance comparisons, despite differences in the size of genes [6, 7]. Development of such tests is difficult, however, as markers are usually correlated with one another and have highly variable minor allele frequencies [15]. As a result, tests have often been born more out of practicality or computational ease. Some gene-based tests take the smallest p-value over all the markers in the region. Others, such as that implemented in PLINK, take a more sophisticated approach, converting p-values of markers to χ^2_1 test statistics, averaging those tests statistics, then comparing it to a null distribution generated from permutations of the outcome under the null [12, 8]. Liu et al. (2010) use a similar, though more efficient method, in which they again convert marker p-values to χ^2_1 test statistics, take the sum of those test statistics, then generate a null distribution by sampling from sums of correlated χ^2_1 random variables. Both approaches are intuitive and valuable ways to assess gene significance, though in both cases the power for detection of a gene becomes not only a function of the effect size of the individual markers, but the degree to which markers are in LD with one another. For example, assuming only one marker in the region has a truly non-zero effect size, power for detecting that effect will be higher if the marker is in high LD with other markers than if it is independent of them. Moskvina et al. (2012) make this same observation, having noticed that the significance of regions they tested changed according to how much they pruned markers in high LD with one another [10].

One way to think about why this phenomenon occurs is that, rather than transforming the test statistic so that markers highly correlated with one another "mean less" because they do not contribute independent information, they transform the null distribution for certain markers under the null to "mean more." As a result, the type 1 error is maintained, but power varies as a function of the correlation between the marker and surrounding markers. Intuitively, this is not a desirable property because it will lead to a systematic under-detection of those loci that happen to be independent of proximal markers even though they are inherently no less important in predicting the outcome. This issue becomes particularly problematic for sequence data since there would generally be even more correlation. However, the issue is a zero-sum trade-off; what results in less power for detection of single nucleotide polymorphisms (SNPs) in low LD translates to more power for detection of SNPs in high LD. Though, if there is an underlying common function or characteristic of those genomic regions whose significant SNPs are not in high LD, perhaps due to when

they first occurred in evolutionary history, such regions will likely be missed in association analyses so that potentially key regions will not be studied in greater depth. As a result of this shortcoming, which may be more or less important depending on the specific LD structure of the genomic region under study, we propose two methods, one of which transforms summary Z-statistics from univariate regressions of markers so that it follows a standard parametric distribution under the null hypothesis and power does not vary with the LD structure, and the second of which uses an eigendecomposition of the information matrix to find the "effective" amount of information in the region and increases power by performing a more parsimonious test. If the information matrix is evaluated under the null, this latter test is essentially a dimension-reduced score test analogue to a method described in [3, 14], which finds the principal components of the data matrix. Specifically, for the first approach we propose, we find Z-statistics associated with each marker in our region and the correlation matrix of the markers and perform a χ^2 test, an approach similar to that proposed by Yang et al. (2012). In case the correlation matrix is imperfectly known, we propose a modification of this test that adjusts the correlation structure to protect the type 1 error. In the latter approach, we calculate the eigenvectors associated with the information matrix to obtain a most powerful linear combination of the scores, on which we again perform a χ^2 test after having normalized by the variance of the loadings. Moskvina et al. (2012) also propose solutions, one of which is based on Hotelling's T^2 test, while another is based on multivariate logistic regression, though concludes that both perform similarly. We compare these various approaches under different structures of LD and effect size. We apply our methods to a case-control sequence data set of oral cleft and an already-published GWAS study of Chronic Obstructive Pulmonary Disorder (COPD) [11].

2 Methods

2.1 Description of permutation tests

First we show how power differs for permutation-based gene tests as a function of linkage disequilibrium from a theoretical perspective. When we refer to permutation-based gene tests, we mean gene-based tests in which the sum of the χ^2 statistics for markers is taken and then an empirical p-value is calculated by permuting case-control status to generate a null distribution. By Imhof (1961), in connection with Liu et al. (2010), we know that the null distribution of the permutation-based test is $\sum_{i=1}^{q} \lambda_i \chi_1^2$, where χ_1^2 is a chi-squared random variable on 1 d.f., λ_i is the i^{th} eigenvalue of Σ, the $q \times q$ correlation matrix of the SNPs comprising the gene to be tested. Under the alternative, the distribution is approximately $\sum_{i=1}^{q} \lambda_i \chi_1^2(\delta_i^2)$, where the non-centrality parameter δ_i is calculated $\delta_i = v_i^t \cdot \mu/\sqrt{\lambda_i}$, where v_i is the eigen vector of Σ corresponding to λ_i, and μ is the q-dimensional mean of the multivariate normal distribution of Z-statistics calculated

for univariate regressions of each SNP. μ is a function of the power for detection of each SNP in the gene. The distribution under the alternative is approximately $\sum_{i=1}^{q} \lambda_i \chi_1^2(\delta_i^2)$, rather than exactly, because the correlation matrix of marker Z-statistics coming from univariate regressions diverges from the correlation structure of the covariates when under the alternative. However, so long as there is not significant variation between observations in the true probability of being a case, this divergence will not be relevant. Since the relative risk of disease conferred by most minor alleles is small, it is likely that the approximation is valid in most studies.

Suppose there is a single causal SNP X_1 and, without loss of generality, it is the first entry in the $(q+2)$-marker gene and q other SNPs, Z_i, $1 \leq i \leq q$, are correlated with it but do not cause the outcome. Also assume that the correlation coefficient between X_1 and Z_i is ρ_i, and the last SNP, X_2, is uncorrelated with X_1 and does not cause the outcome. The first entry of μ, which represents the mean of the Z-statistic for X_1, can be written $k_1 \cdot \sqrt{n}$ for some number k_1 where n is the sample size. Since the asymptotic relative efficiency for using Z_i rather than X_1 is ρ_i^2, the i^{th} entry of μ, that associated with Z_i, can be written $k_1 \cdot \rho_i \cdot \sqrt{n}$ [13, 5]. The entries of μ_1, the $(q+2)$-dimensional mean of the Z-statistics corresponding to the permutation-based gene test when the causal SNP is X_1, are

$$\mu_{\mathbf{1}}{}^T = (k_1\sqrt{n},\ k_1 \cdot \rho_1\sqrt{n},\ k_1 \cdot \rho_2\sqrt{n}, \ldots,\ k_1 \cdot \rho_q\sqrt{n},\ 0).$$

In contrast, suppose that the correlation structure among SNPs is the same, that is, $\mathrm{Cor}(X_1, Z_i) = \rho_i$ for $1 \leq i \leq q$, but X_1 does not cause the outcome, and instead X_2, the SNP uncorrelated with all other $(q+1)$ markers, causes the outcome and to the same degree as X_1 did so in the previous scenario. Then μ_2, the $(q+2)$-dimensional mean in this case is

$$\mu_{\mathbf{2}}{}^T = (0,\ 0,\ 0, \ldots,\ 0,\ k_1\sqrt{n}).$$

If $Q \equiv \sum_{i=1}^{q} \lambda_i \chi_1^2(\delta_i^2)$, the power of an size α (i.e., type 1 error rate) permutation-based gene test is $P(Q > c^*)$, where c^* is the $(1-\alpha)$ quantile of the random variable $\sum_{i=1}^{q} \lambda_i \chi_1^2$. The intuition behind the power gains for causal SNPs in regions of LD is that the non-centrality parameters δ_i will generally be larger when the causal SNP is in a region of LD than when it is not. Providing greater rigor than this intuition is difficult because the calculation of δ_i for all $1 \leq i \leq (q+2)$, even in the simple case of $\rho_i = \rho_j$ for all $1 \leq i \leq q$, can be complicated. However, sampling from the appropriate distributions demonstrates that there is greater power to detect a gene-outcome association when the causal SNP is in a region of LD. Figure 1 shows that under the alternative of gene-outcome association and for a fixed effect size and $\rho_i = \rho_j = \rho$

for $1 \leq i \leq q$, i.e., the $(q+2) \times (q+2)$ correlation matrix $\mathbf{\Sigma}$ is

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & \rho & \rho & \dots & \rho & 0 \\ \rho & 1 & \rho & \dots & \rho & 0 \\ \rho & \rho & 1 & \dots & \rho & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho & \rho & \rho & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix},$$

the distribution of the test statistic Q when the causal SNP in the gene is in a region of LD is stochastically greater than the distribution when the causal SNP in the gene is not in a region of LD. As a result, power is greater to detect a gene whose causal SNP is in a region of LD for a test of any size α in a permutation-based gene test. Figure 1 was generated assuming a gene consisting of 7 markers, where 6 markers were correlated with coefficient ρ, shown for values 0.2, 0.5, and 0.8 in the figure, with the causal SNP in the LD block (y-axis) versus not in the LD block (x-axis). While the example may seem contrived, if we consider $q = 10$ so that our gene consists of 12 SNPs in total, the correlation structure in this example is similar to choosing the first 12 SNPs of BRCA1 [2], in which case $\rho \approx 0.96$, and where the last row and column above would be approximately 0.24 rather than 0.

Figures 2 and 3 demonstrate from a graphical perspective how permutation-based gene tests can have variable power as the LD structure changes. The figures illustrate that there is more power for the permutation-based gene test when causal SNPs are in high LD blocks as compared to causal SNPs in low LD blocks. Additionally, if a causal SNP is not in LD with other SNPs, but large LD blocks exist in the gene, power for the permutation-based gene test decreases as the size of the block increases. Data were generated with a minor allele frequency of all SNPs of approximately 0.3 with Hardy-Weinberg equilibrium assumed, and, within the LD block, correlation between SNPs was approximately 0.65, whereas SNPs not in the LD block were independent of one another. The gene consisted of 20 SNPs, and there were 600 subjects with an equal number of cases and controls. Power calculations were based on 1000 iterations at each effect size (Figure 2) or LD block size (Figure 3). We calculated power at 18 different effect sizes (Figure 2), with the effect size ranging from a log OR of 0 to 1.2, and 20 different LD block sizes (Figure 3), with the LD block size ranging from 1 SNP to 20 SNPs. So when the size was 20 SNPs, the LD block was the entirety of our hypothetical gene (Figure 3). Binary outcomes were generated assuming a logistic regression model.

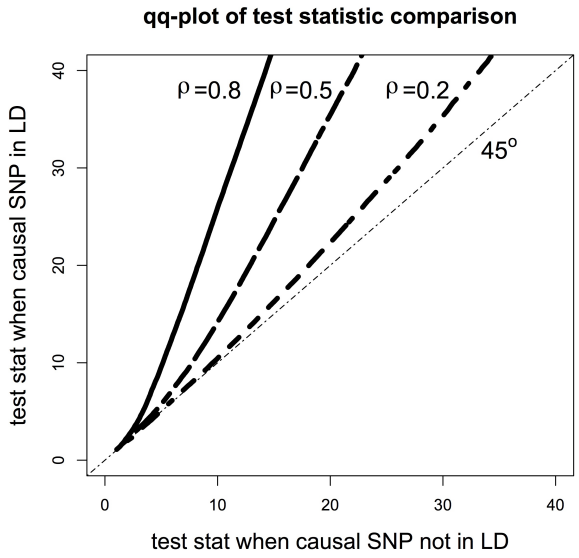

Figure 1: The higher the correlation in the LD block containing the causal SNP, the more power relative to the causal SNP not in the LD block using the permutation-based test.

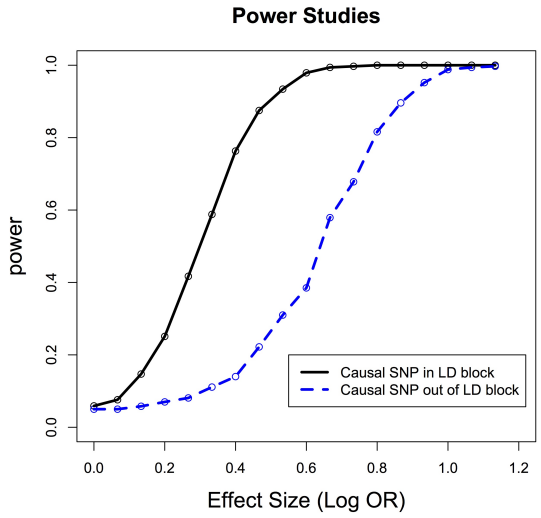

Figure 2: There is more power to detect a SNP in high LD with other, non-causal SNPs, than a SNP in low LD for an identical effect size.

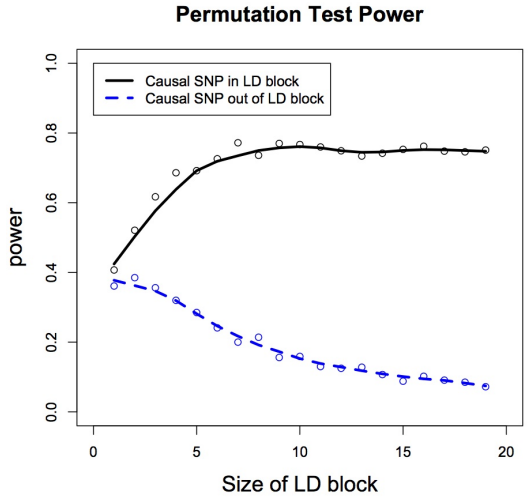

Figure 3: For a constant effect size, size of the LD block in which the causal SNP is located is related to the power for detection.

3 Summary statistic based test

We first describe a simple solution to the problem of how LD structure can affect the power to detect genomic regions in which there are significant SNPs. Our solution is based on the Z-statistics associated with each marker and the correlation matrix of the SNPs. Since we propose this test as one that can be used without a full data set, we propose a modification of it in case the true correlation structure is not perfectly known

or it is believed that study participants are not reflective of the population from which the correlations of SNPs are calculated (such as with HapMap reference panels).

3.1 Description

One solution to the problem of under-detection of SNPs in low LD posed by permutation tests is transformation of the gene-based test statistic so that under the null it follows a standard parametric distribution, rather than creating a non-standard null distribution through permutations. One way to accomplish this task, and one in which it is unnecessary to reanalyze data, is to perform a joint test on the Z-statistics coming from a univariate regression model for each marker. It is a an approach similar to that described by Yang et al. (2012), though uses summary statistics directly rather than estimated model coefficients. Since the estimated covariance structure of these statistics under the null is the correlation of the markers themselves [1], one can use the data to estimate the covariation of the Z-statistics or an online database of LD or correlation structure of SNPs. We show the result in the context of a logistic regression model as is generally used in case-control studies, though the result is identical for other parametric models. The intuition behind this result is that if two markers are highly correlated, then when by chance under the null, one marker is significant (or insignificant), the other marker will similarly be significant (or insignificant). However, if two markers are not correlated, then the chance significance or insignificance of one marker will not inform the significance of the other marker. And since Z-statistics have variance 1 by definition, their covariance matrix is identical to their correlation matrix. Thus, supposing we have q markers, which, from previous studies, are known to have Z-statistics of $\mathbf{Z} = (Z_1, \ldots, Z_q)^T$, and which have correlation structure $\mathbf{V}$, then under the null hypothesis of no marker being associated with the outcome, $T \equiv \mathbf{Z}^T V^{-1} \mathbf{Z} \sim X^2_q$. One then rejects the null of no association between the region composing the q markers and the outcome for an extreme value of the test statistic T using a pre-determined α level.

If one is not confident that V accurately reflects the covariance of the SNPs in the data matrix and therefore $\mathbf{Z}$ under the null hypothesis, it is possible to construct a more conservative test by shrinking the off-diagonal elements of V towards 0. Thus, if V is an estimate of the covariance of the SNPs, one can compute $V^*_\gamma \equiv \gamma V + (1-\gamma)\mathbf{I}_q$, where $\mathbf{I}_q$ is a $q \times q$ identity matrix and $0 \leq \gamma \leq 1$.

Again using [4], if $Z \sim MVN(0, V)$ but we use V^*_γ as an estimate of the correlation structure in the gene based test, then $Z^T V^{*-1}_\gamma Z \sim \sum_{i=1}^q \lambda_i \cdot \chi^2_1$ where q is the dimension of vector Z, and where λ_i is the i^{th} eigenvalue of $\mathbf{V}\mathbf{V}^{*-\mathbf{1}}_\gamma$. By construction of V^{*-1}_γ, $\sum_{i=1}^q \lambda_i < \dim(Z)$ for $0 < \gamma < 1$, where $\dim(\cdot)$ the dimension of the vector argument. This fact in itself does not not necessarily imply a more conservative test for all size α tests because when eigen values are not equal to one another as is the case for the decomposition of $\mathbf{V}\mathbf{V}^{*-\mathbf{1}}_\gamma$ with $\mathbf{V} \neq \mathbf{V}^{*-\mathbf{1}}_\gamma$, $\sum_{i=1}^k \lambda_i < \dim(Z)$ can be true, but $Z^T \mathbf{V}^{*-\mathbf{1}}_\gamma Z$ is not stochastically less than

$\chi^2_{\dim(Z)}$, the null distribution of the test statistic when the correlation structure is correctly known. However, for modest values of γ (i.e., 0.8-1.0, where 1.0 corresponds to no transformation), the test using the adjusted correlation matrix will generally be more conservative. It is difficult to obtain simple solutions for how much conservative a test will be using this modification since it will depend on the quantile corresponding to the intended type 1 error and the specific $\mathbf{V}\mathbf{V}^{*-\mathbf{1}}_{\gamma}$. Thus, to give a practical sense of useful values of γ, we borrowed a correlation structure of SNPs in the INS-IGF2 gene of Chromosome 11 from the CEU reference panel in one case and the CHB+JPT reference panel in the other case [2]. If the true, underlying population giving rise to the SNPs was more reflective of the CEU reference panel, but the analyst incorrectly guessed the correlation structure to be that of the CHB+JPT reference panel when performing the summary statistic gene-based test,the type 1 error rate for a nominal 0.05 size test would in fact be a highly inflated 0.61. Similarly, if the true, underlying population giving rise to the SNPs was more reflective of the CHB+JPT reference panel, but the analyst incorrectly guessed the correlation structure to be that of the CEU reference panel for the summary statistic gene-based test, the type 1 error rate for a nominal 0.05 size test would be 0.69. If the type 1 error is inflated in one scenario, there is no implication that it will be deflated in the 'inverse' scenario.

In the scenario where the underlying population was more reflective of the CEU panel, the type 1 error using our modified summary statistic test with adjusted correlation matrix and γ's of 0.9, 0.5, and 0.3 led to reduced error rates of 0.36, 0.11, and 0.09, respectively, instead of 0.61. When the underlying population was CHB+JPT, the type 1 error using our adjustment correlation matrix and γ's of 0.9, 0.5, and 0.3 led to reduced error rates of 0.45, 0.10, and 0.07, respectively, instead of 0.69. While in all cases, the nominal size of the test is not quite achieved, type 1 error is greatly reduced, and in some cases will be achieved when divergence between correlation structures of the true and hypothesized populations are not as great as that in these scenarios. The greatest reduction in type 1 error occurs with initial deviation of γ from 1; i.e., a movement of γ from 1 (indicating an unadjusted correlation matrix) to 0.9 will reduce type 1 error more than a movement of 0.6 to 0.5. And, as mentioned, with very small values of γ, there is not necessarily a guarantee of continued reduction in type 1 error for some nominal α level tests, nor should such γ values be used if indicative of no confidence in one's estimated correlation matrix.

To simulate a less drastic divergence between true and estimated correlation matrices and assess error rates and the proposed adjustment method in that context, we generated correlation matrices whose entries were beta-distributed random variables with means corresponding to the entries in the CHB+JPT reference panel and standard deviation approximately 0.03-0.04 (approximately because standard deviation is partly a function of the mean). With a population whose underlying correlation structure was in truth reflective of the CHB+JPT panel, but using the generated correlation matrices in our calculations of the test statistic, the

average type 1 error rate was 0.19. Adjusting the generated correlation matrices according to our method and with a γ value of 0.95, the error rate was reduced to 0.05. Adjustment of the generated correlation matrices with a γ value of 0.90 led to a type 1 error rate of 0.03.

The summary statistic based test we have proposed is a viable way of performing gene-based testing when one does not want power to vary as a function of the correlation structure of the SNPs composing the gene. A weakness of such an approach is an inability to know the underlying correlation structure of the SNPs used in the unvariate regression analyses givingn rise to the Z-statistics used in the summary statistic test. We have shown that incorrect guesses of the underlying correlation structure can lead to a significant increase in the type 1 error rate and therefore have proposed an adjustment method which can lead to achievement of error rates in line with the nominal size of the test. However, since by supposition of this setting the correlation structure of SNPs is never known, it is impossible to know the needed value of γ. As a result, it may be best to perform one's summary statistic based test with γ values ranging from 0.8-1.0 as a sensitivity analysis to see how one's conclusions change based on different values. Values of γ smaller than 0.8 probably reflect little confidence in the estimated correlation structure, in which case feasibility of the analysis in the first place should be reassessed.

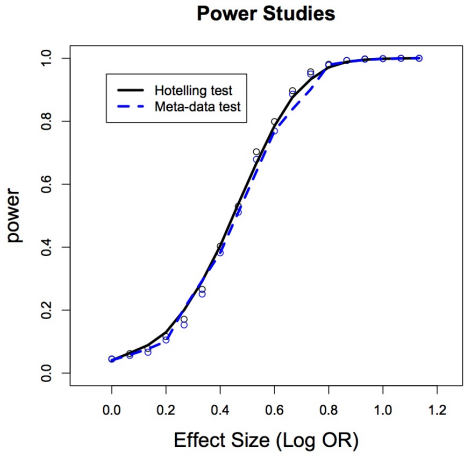

Figure 4: Power comparison of Moskvina et al.'s method with our summary statistic based method when the causal SNP is in a block of high LD.

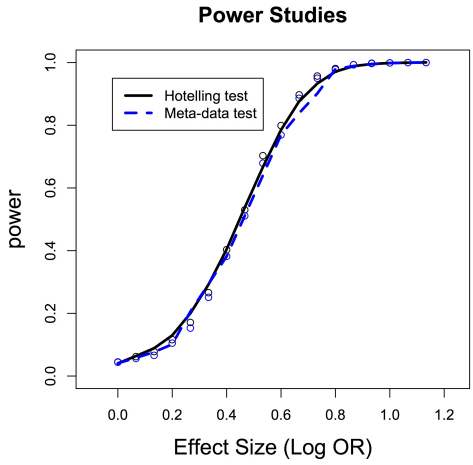

Figure 5: Power comparison of Moskvina et al.'s method with our summary statistic based method when the causal SNP is not in a block of high LD, but there is an LD block elsewhere in the gene.

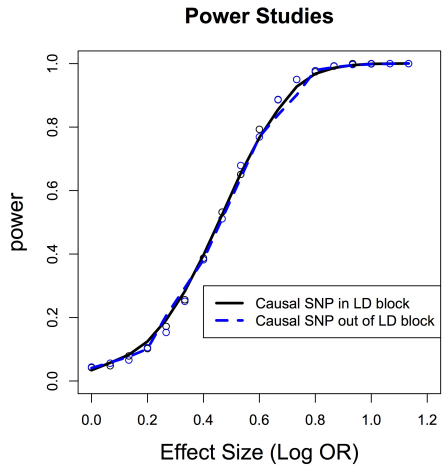

Figure 6: A SNP in high LD with other, non-causal SNPs, has no more power to be detected with the summary statistic test than a SNP in low LD, as desired.

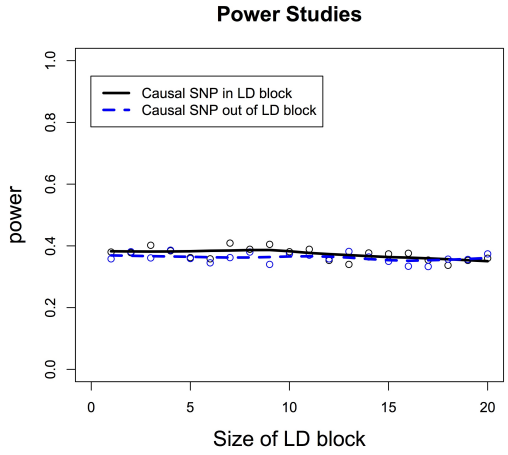

Figure 7: Test working as desired since power is constant across LD block sizes for fixed effect size.

4 Eigen decomposition-based test

4.1 Description

The above approach controls for the LD structure of the region under study by transforming the test statistic so that LD no longer affects the power to detect significant regions. However, there are other ways one can make use of the LD structure to construct more powerful tests, such as by dimension reduction. Consider an extreme example where an investigator is interested in a region with d SNPs, and these SNPs are in nearly perfect LD so that a correlation matrix of them has off-diagonal entries close to 1. Because they are highly correlated, the association between any SNP and the outcome adds little information on top of that between any other SNP and the outcome. As a result, intuition may tell us that using a d-d.f. test on the region after having properly accounted for the underlying LD structure is not the most powerful approach since there is essentially the information of 1 SNP contained in the entire region. On the other hand, it is difficult to justify focusing on any one SNP over another as one might do when "tagging" the region. Also, while no additional SNP contributes much information over another, there is still some amount of additional information contained in each one that, ideally, would not be ignored.

Finding the eigenvectors and values of the information matrix is one way to approach this scenario. It gleans the essential information from the LD block, thus stripping away extraneous information that dilutes the power of proposed tests while avoiding the arbitrariness of pruning the number of SNPs being examined. It is an approach similar to finding the principal components of the data matrix and then regressing the outcome on those components if the information matrix is evaluated under the null [3, 14] and may even be

thought of as a score test analogue to it. If certain covariates have been shown to control for population stratification, it also may be fitting for the matrix to be evaluated under the alternative using the estimated effect sizes of those covariates. Also, as simulations demonstrate, there may be power gains under certain correlation structures or when effect or sample sizes are small. Since the information matrix is the covariation of the scores associated with each marker and since score functions of highly correlated markers are correlated as well, identifying the chief axes of the covarying scores is synonymous with finding the eigenvectors of the information matrix. One can then detect small deviations from the mean under the null hypothesis, the $\mathbf{0}$ vector, by performing a parsimonious test. Additionally, if we are considering the underlying model to be that of logistic regression, both the information matrix and the score have simple forms and are computationally tractable.

We now describe how to construct the test, which will place no constraint on estimation of the intercept, but reduce the dimension of the covariates underlying the scores. The score function associated with the j^{th} marker under a logistic regression model is

$$S(\beta_j) = \frac{\partial L(\beta)}{\partial \beta_j} = \sum_i y_i x_{ij} - \sum_i n_i x_{ij} \frac{\exp(\sum_k \beta_k x_{ik})}{1 + \exp(\sum_k \beta_k x_{ik})}.$$

While the $(j,l)^{th}$ entry in the information matrix is

$$\begin{aligned} I_{q+1\ (j,l)} &= Cov(S(\beta_j), S(\beta_l)) = Cov(\sum_i y_i x_{ij}, \sum_i y_i x_{il}) \\ &= \sum_i x_{ij} x_{il} P(y_i = 1)\{1 - P(y_i = 1)\} = \sum_i x_{ij} x_{il} \frac{\exp(\sum_{m=0}^{k} \beta_m x_{im})}{\{1 + \exp(\sum_{m=0}^{k} \beta_m x_{im})\}^2}, \end{aligned}$$

where we could estimate $P(y_i = 1)$ under an intercept-only model, i.e., the proportion of cases, or another model that included potential confounders. For the sake of explanation, we will proceed as if using the intercept-only model. The information matrix for the logistic regression model is then

$$\mathbf{I_{q+1}} = \begin{bmatrix} I_{q+1\,(1,1)} & I_{q+1\,(1,2)} & \cdots & I_{q+1\,(1,q+1)} \\ I_{q+1\,(2,1)} & I_{q+1\,(2,2)} & \cdots & I_{q+1\,(1,q+1)} \\ \vdots & \vdots & \ddots & \vdots \\ I_{q+1\,(q+1,1)} & I_{q+1\,(q+1,2)} & \cdots & I_{q+1\,(q+1,q+1)} \end{bmatrix},$$

where, as is consistent from our definition of $I_{j,l}$, $I_{l,j} = I_{j,l}$. Also, I is $(q+1) \times (q+1)$ because there are q markers and 1 intercept available for use in the model.

Now define $\mathbf{I_q}$ to be the $q \times q$ information matrix for the covariates, not including the intercept. That

is, $\mathbf{I_q}$ is $\mathbf{I_{q+1}}$ without the first column and first row of $\mathbf{I_{q+1}}$. $\mathbf{I_q}$ can be decomposed into $\mathbf{E\,W\,E^{-1}}$, where $\mathbf{E} \equiv (\mathbf{e_1}, \mathbf{e_2}, \cdots, \mathbf{e_q})$ is a matrix of the q eigenvectors $\mathbf{e_i}$, $1 \leq i \leq q$, and $\mathbf{W} \equiv \text{diag}(\lambda_1, \lambda_2 \cdots, \lambda_q)$ is a diagonal matrix of the corresponding eigenvalues λ_i, $1 \leq i \leq q$. While one can use a variable number of eigenvectors in the analysis, if we suppose that we are in the situation described above where all d markers are highly correlated, then making use of just the first component may be sufficient to adequately encompass the information contained in the genomic region. More generally, a systematic criterion for deciding which eigen vectors to use is employing all those whose associated eigenvalues are larger than the average eigenvalue.

For the sake of explanation, we suppose first that we will construct the test using only the eigenvector associated withe largest eigenvalue and then generalize later. Denote $\mathbf{e_1}$ the first column of $\mathbf{E}$ and vector associated with the largest eigenvalue (assume the columns of $\mathbf{E}$ are ordered according to decreasing eigenvalue). The interpretation of $\mathbf{e_1}$ is the axis of maximum variation of the distribution whose covariance matrix is $\mathbf{I_q}$, and λ_1, the associated eigenvalue, can be interpreted as the variation along that axis. Since $\mathbf{I_q}$ is $q \times q$, $\mathbf{e_1}$ is a $(q \times 1)$ unit eigen vector. Define a new 2×2 information matrix $\mathbf{I}^*$ as ${\mathbf{I}^*}_{(1,1)} \equiv \mathbf{I}_{q+1\,(1,1)}$, ${\mathbf{I}^*}_{(2,2)} \equiv \lambda_1$, and ${\mathbf{I}^*}_{2,1} = {\mathbf{I}^*}_{1,2} = \mathbf{e_1}^T \cdot \mathbf{I_{q\,(,1)}}$, where $\mathbf{I_{q\,(,1)}}$ is the first column of $\mathbf{I_q}$, $\mathbf{v}^T$ denotes the transpose of vector $\mathbf{v}$, and $\mathbf{e_1}^T \cdot \mathbf{I_{q\,(,1)}}$ denotes the dot product of vectors $\mathbf{e_1}$ and $\mathbf{I_{q\,(,1)}}$. The test statistic for the 1 d.f. score test analogue of the method described in [3, 14] is then $(\mathbf{S}^t \cdot \mathbf{e_1})^2 \cdot [(\mathbf{I}^{*\,-1})_{(2,2)}]$, where $\mathbf{S}$ is the $q-$dimensional vector of scores associated with the q markers, which follows a χ^2_1 distribution under the null hypothesis of no gene-outcome association.

To generalize the method to using p eigenvectors, similar to regressing the outcome on the first p principal components of the data matrix, again perform an eigendecomposition of $\mathbf{I_q}$, and define $\mathbf{e_1}, \ldots, \mathbf{e_p}$ as the p unit eigenvectors of length q associated with the p largest eigenvalues. Call those associated eigenvalues $\lambda_1, \ldots, \lambda_p$. Define a new information matrix $\mathbf{I}^{**}$ as ${\mathbf{I}^{**}}_{(1,1)} = \mathbf{I_{q+1\,(1,1)}}$, ${\mathbf{I}^{**}}_{(m,m)} = \lambda_m$ (where $2 \leq m \leq p$), ${\mathbf{I}^{**}}_{(m,n)} = {\mathbf{I}^{**}}_{(n,m)} = 0$ (where $m \neq n$ and $2 \leq m, n \leq p$), and ${\mathbf{I}^{**}}_{(m,1)} = {\mathbf{I}^{**}}_{(1,m)} = \mathbf{e_m}^T \cdot \mathbf{I_{q\,(,1)}}$ (where $1 < m \leq p$ and $\mathbf{I_{q\,(,1)}}$ denotes the first column of $\mathbf{I_q}$). Note that the off-diagonals of $\mathbf{I}^{**}$ which are neither the first row nor first column are zero by the orthogonality of eigenvectors; i.e., for $1 < m, n \leq p$, $m \neq n$, ${\mathbf{I}^{**}}_{(m,n)} = {\mathbf{I}^{**}}_{(n,m)} = \mathbf{e}_n^T \cdot \mathbf{I_q} \cdot \mathbf{e_m} = \mathbf{e}_m^T \cdot \mathbf{I_q} \cdot \mathbf{e_n} = 0$. $\mathbf{I}^{**}$ is $(p+1) \times (p+1)$ and looks as follows

$$\mathbf{I}^{**} = \begin{bmatrix} I_{q+1\,(1,1)} & \mathbf{e_1}^T \cdot \mathbf{I_{q\,(,1)}} & \mathbf{e_2}^T \cdot \mathbf{I_{q\,(,1)}} & \ldots & \mathbf{e_p}^T \cdot \mathbf{I_{q\,(,1)}} \\ \mathbf{e_1}^T \cdot \mathbf{I_{q\,(,1)}} & \lambda_1 & 0 & \ldots & 0 \\ \mathbf{e_2}^T \cdot \mathbf{I_{q\,(,1)}} & 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{e_p}^T \cdot \mathbf{I_{q\,(,1)}} & 0 & 0 & \ldots & \lambda_p \end{bmatrix}$$

Define ${\mathbf{I}^{**}}^{-1}_{p\times p}$ as the lower-right $p \times p$ sub-matrix of ${\mathbf{I}^{**}}^{-1}$. The test statistic is $\mathbf{S}^T \cdot (\mathbf{e_1} \ \ldots \ \mathbf{e_p}) \cdot {\mathbf{I}^{**}}^{-1}_{p\times p} \cdot (\mathbf{e_1} \ \ldots \ \mathbf{e_p})^T \cdot \mathbf{S}$, which follows a χ^2_p distribution under the null hypothesis of no gene-outcome association, where again $\mathbf{S}$ is a vector of scores of length q.

Oftentimes in GWAS, population stratification can obscure the relationship between markers (or groups of markers) and outcomes. In these settings, it is necessary to account for stratification by fitting models with covariates or ancestry informative markers (AIM) that adjust for the different populations composing the sample. Reducing the dimension of such covariates along with the markers making up the gene renders them less effective if not useless for their intended purpose of controlling for population stratification. Thus, it is necessary to construct a score test where only a chosen subset of the covariates have their dimension reduced and the information matrix is evaluated under the alternative for those covariates whose dimension is not reduced. Doing so is not difficult and only requires treatment of the adjusting covariate in the quasi-information matrix as we treated the intercept in $\mathbf{I}^{**}$, where the off-diagonal entries were a linear combination of the appropriate eigen vector and $q-$length sub-column of the original information matrix. So suppose there are q markers and we only want to reduce the dimension of the last $(q-1)$ of this group. Let $\mathbf{I_{(q+1)}}$ be the $(q+1)\times(q+1)$ information matrix and define $\mathbf{I_{(q-1)}}$ as the lower right $(q-1)\times(q-1)$ sub-matrix of $\mathbf{I_{(q+1)}}$. Decompose $\mathbf{I_{(q-1)}}$ into $\mathbf{E'\,W'\,E'^{-1}}$, where $\mathbf{E'}$ is the matrix of $(q-1)$ eigenvectors, e'_j for $1 \le j \le (q-1)$, of $\mathbf{I_{(q-1)}}$ and $\mathbf{W'}$ is the diagonal matrix of corresponding eigenvalues, λ'_j for $1 \le j \le (q-1)$, and we use $\mathbf{E'}$ and $\mathbf{W'}$ to differentiate these matrices from those defined above and *not* to indicate the transpose of these matrices. Suppose we want to use the first p' eigenvectors for our test of the $(q-1)$ markers in the group whose dimension we reduce and where $p' \le (q-1)$. The quasi-information matrix is defined

$$\mathbf{I}^{***} = \begin{bmatrix} I_{q+1\,(1,1)} & I_{q+1\,(1,2)} & \mathbf{e'_1}^T \cdot \mathbf{I_{q-1}}_{(,1)} & \ldots & \mathbf{e'_{p'}}^T \cdot \mathbf{I_{q-1}}_{(,1)} \\ I_{q+1\,(2,1)} & I_{q+1\,(2,2)} & \mathbf{e'_1}^T \cdot \mathbf{I_{q-1}}_{(,2)} & \ldots & \mathbf{e'_{p'}}^T \cdot \mathbf{I_{q-1}}_{(,2)} \\ \mathbf{e'_1}^T \cdot \mathbf{I_{q-1}}_{(,1)} & \mathbf{e'_1}^T \cdot \mathbf{I_{q-1}}_{(,2)} & \lambda'_1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{e'_{p'}}^T \cdot \mathbf{I_{q-1}}_{(,1)} & \mathbf{e'_{p'}}^T \cdot \mathbf{I_{q-1}}_{(,2)} & 0 & \ldots & \lambda'_{p'} \end{bmatrix}.$$

Analogous to the test statistic defined in the previous test, define ${\mathbf{I}^{***}}^{-1}_{p'\times p'}$ as the lower-right $(p' \times p')$ sub-matrix of ${\mathbf{I}^{***}}^{-1}$ and $\mathbf{S'}$ as the vector of scores associated with the $(q-1)$ markers whose dimension we reduce. The test statistic is $\mathbf{S'}^T \cdot (\mathbf{e'_1} \ \ldots \ \mathbf{e'_{p'}}) \cdot {\mathbf{I}^{***}}^{-1}_{p'\times p'} \cdot (\mathbf{e'_1} \ \ldots \ \mathbf{e'_{p'}})^T \cdot \mathbf{S'}$, which follows a $\chi^2_{p'}$ distribution under the null hypothesis of no gene-outcome association.

5 Simulation results

5.1 Summary statistic based test and comparison with Hotelling's T^2

Moskvina et al. propose a test based on Hotelling's T^2. If one knows the true information matrix, it is a multivariate score test and follows a X^2 distribution under the null. Supposing that there are q markers and $\mathbf{S} = (S(\beta_1), \ldots, S(\beta_q))^T$, the associated scores, and the true information matrix is $\mathbf{I}$, then under the null of no marker being associated with the outcome, $\mathbf{S}^T\mathbf{I}^{-1}\mathbf{S} \sim X^2_q$. Similarly, and as described above, the summary statistic based test uses the Z-statistics associated with univariate logistic regression models, $\mathbf{Z}$, and the marker correlation matrix, $\mathbf{V}$, so that under the null hypothesis and assuming V is perfectly known, $T \equiv \mathbf{Z}^T V^{-1}\mathbf{Z} \sim X^2_q$. While both of these approaches use similar information (i.e., some measure of SNP significance not controlling for other SNPs and an estimate relating to the correlation of those measures), in simulation the summary statistic based approach seems to have slightly less power than the Hotelling's T^2 test, but the difference is almost non-existent in many cases (Figures 4 and 5), and the summary statistic based test also seems to be more conservative, again assuming a perfectly known correlation structure V. More importantly, however, the summary statistic approach does not require individual-level data, which is not the case with Hotelling's T^2. Power for the summary statistic based approach does not vary as a function of whether the causal SNP in the gene is in a region of LD or not (Figure 6), and, for a constant effect size, power does not vary as a function of the size of the LD block in the gene (Figure 7).

Simulations were generated under the same framework as we used with the permutation test simulation above. Covariates were generated with a minor allele frequency of approximately 0.3, and, within any LD block, correlation between SNPs was again approximately 0.65, whereas SNPs not in the LD block were independent of one another. We assumed Hardy-Weinberg equilibrium. The gene consisted of 20 SNPs, and there were 600 subjects with an equal number of cases and controls. Power calculations were based on 1000 iterations at each effect size (e.g., Figure 4) or LD block size (e.g., Figure 5). Lastly, binary outcomes were generated assuming a logistic regression model.

5.2 Eigen decomposition-based test

We examine in simulation performance of the dimension-reduced score test when a single causal SNP was in an LD block and compare this proposed test with the method described in [3, 14], in addition to 1 d.f. score and Wald tests of the causal SNP and a 1 d.f. Wald test of a tagging SNP when the correlation between tagging and causal SNPs was approximately 0.8. Figure 8 shows relative performance of these methods when there was no LD, while Figure 9 compares methods when correlation was approximately 0.15. We see that the performance of the Eigen decomposition-based test performed better relative to the method proposed

in [3, 14] when the LD block was more weakly correlated. As the correlation increases, power of these two methods converges. Direct testing of the causal SNP, be it through a Wald or score test, performed best as expected, though of course knowledge of the true causal SNP is generally never known. Thus, we note that the eigen decomposition-based test performs better than testing of a tagging SNP and makes unnecessary the need to decide which tagging SNP to use. In Figure 9, even under weak LD, the eigen decomposition-based test pays little price in terms of power for no knowledge of the true causal SNP.

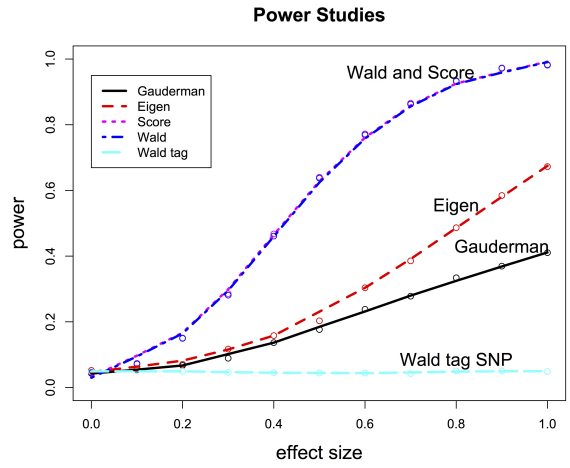

Figure 8: Power comparison between the Eigen-based test and Gauderman's method, along with a direct test of the causal SNPs and tagging SNP under no LD.

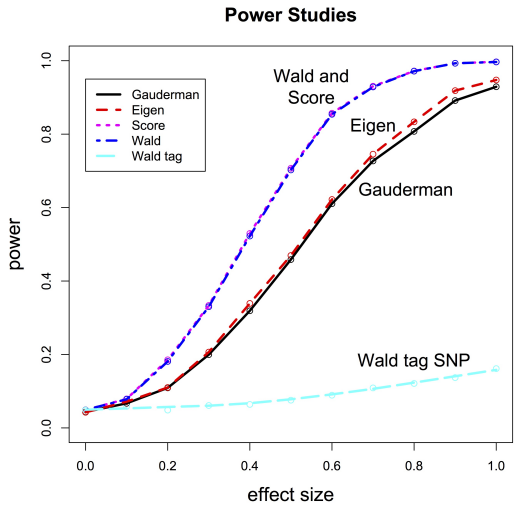

Figure 9: Power comparison between the Eigen-based test and Gauderman's method, along with a direct test of the causal SNPs and tagging SNP under high LD.

6 Data analysis

We analyze a sequence data set composed of 192 cases of cleft lip and 192 controls, on whom we have data for 14 SNPs. The data come from a GWAS in which a candidate gene was identified and then sequenced [9]. We prune the data set so that any observations with missing values or deletions are excluded, giving 172 cases and 176 controls. We also prune SNPs so that any SNP with a MAF less than 0.02 among either cases or controls is excluded, leaving 8 SNPs. We calculate the correlation matrix of SNPs by pooling cases and controls. Using the summary statistic based test, we find that the region composed of the 8 SNPs is associated with cleft lip (p=0.06). Using the eigen decomposition based test with the two eigenvectors whose associated eigenvalue is bigger than the average eigenvalue, we calculate a p-value of 0.017; using 3 eigenvectors such that more than 80% of the variation in scores is explained, we calculate a p-value of 0.016. Thus, as is consistent with the potential power gains posed by dimension reduction, this latter test shows a stronger association between the region of 8 SNPs and cleft lip. For comparison, we also calculated

a permutation test p-value, giving 0.008 (5000 permutations), and a Hotelling's T^2 p-value, giving 0.056 (non-parametric, permutation-based p-value for this test gives 0.057). Assuming little correlation among SNPs, one would expect the permutation test p-value to give a p-value similar to that of the summary statistic based test. The greater significance of the permutation test p-value suggests that a significant SNP is in LD with other SNPs and examination of the data matrix confirms this idea; a SNP whose p-value is 0.018 using a univariate logistic regression model is highly correlated with one SNP (r=0.71) and moderately correlated with another SNP (r=0.43). Since only 8 SNPs are being analyzed, these two SNPs in LD with the significant SNP may be driving the significance of the permutation test.

We also apply the summary-statistic based test to results borrowed from an already-published GWAS along with information on the correlation of markers taken from HapMap [2, 11]. Pillai et al. (2009) identified 5 SNPs in the CDKAL1 gene on Chromosome 6 to be associated with Chronic Obstructive Pulmonary Disorder (COPD). We run our summary statistic based test on their results. Since the results come from a study of Norwegians, we use the (CEU) reference panel from HapMap as an estimate of the correlation structure of SNPs. The underlying population is unlikely to be identical, however, and so we adjust the correlation matrix, shrinking the off-diagonal elements toward 0 as described in the modification of the summary statistic based test to preserve the type 1 error rate. We do so with γ values of 1 (i.e., assuming the correlation structure is correct), 0.9, and 0.8, and corresponding p-values for the 5 d.f. test are 0.0066, 0.003, and 0.001. While the summary statistics we use are borrowed from the 100 most significant SNPs of their analysis [11], the high level of significance for tests corresponding to all γ values and non-arbitrary choice all SNPs in the chosen gene suggest that there is likely some association between the CDKAL1 gene and COPD. Since the test statistics are themselves random variables, specific realizations of them are not necessarily associated with increasing p-values as one would anticipate with decreases in γ. However, work above has shown that, in general, modest decreases in γ should help preserve type 1 error.

7 Discussion

With the availability of sequence data and GWAS, the importance of statistical analysis is shifting from single-locus tests to multi-loci tests that can cover genomic regions, e.g., genes or even pathways. The motivation for this development is to test a hypothesis more grounded in biology and, at the same time, to reduce the multiple testing problem and allow for many SNPs with a small effect size to increase the power of the test by their combined inclusion in the model. One of the theoretical issues that has so far not been addressed adequately is the impact of LD on the power of the test statistic in permutation-based gene tests. Controlling for the LD between loci is important to assess the relative importance of the different regions

that are tested, especially when LD heterogeneity between regions is significant. In this paper, we have proposed 2 approaches that address this issue.

While our summary statistic based test may give one similar results to a Hotelling's T^2 based test, the summary statistic test does not require the original marker data from which Z-statistics are calculated. This unique advantage opens up the possibility for more in-depth analysis of previously published studies, and, with sufficient methodological development, could even suggest summary statistic based pathway analyses when combined with summary statistics from expression analyses. It also opens up the possibility of cross-study gene-based tests, where Z-statistics from the same markers are combined across previously published GWAS to reap power gains. A shortcoming of our summary statistic based test is that if the estimated correlation structure used in the test is not reflective of the underlying population, the test may suffer from inflated type 1 error. We therefore proposed an modification of the test by adjusting the estimated correlation matrix, which, in general, should help control the error rate. If there is insufficient justification for why the estimated correlation matrix is representative of the underlying population, the test should not be used even with correlation matrix adjustment.

Both of the proposed gene-based tests in this paper fail to describe the direction of association between the gene and outcome, instead describing only significance of association. Direction of association is a difficult concept to interpret when a gene is composed of multiple SNPs, with some alleles protective and others a risk factor for the outcome. One goal in gene-based testing might be to gain an understanding of such a concept. Additionally and with regard to dimension reduction approaches, if alleles in a dimension-reduced block of SNPs are both protective and harmful, there could be a loss of power using a dimension-reduction gene-based test. A test that used a priori analyses to decide whether alleles are protective or harmful and, in turn, used that information to inform the dimension reduction process might be another valuable area of research in gene-based testing.

8 Acknowledgements

The authors wish to thank Dr. Xihong Lin in preparation of this manuscript. This work was supported by the National Institutes of Health [Training Grant T32 NS048005]. *Conflict of Interest*: None declared.

References

[1] Karen N Conneely and Michael Boehnke. So many correlated tests, so little time! rapid adjustment of p-values for multiple correlated tests. *The American Journal of Human Genetics*, 81(6):1158–1168, 2007.

[2] The International HapMap Consortium. Computing the distribution of quadratic forms in normal variables. *Nature*, 426:789–796, 2003.

[3] W Gauderman, Cassandra Murcray, Frank Gilliland, and David Conti. Testing association between disease and multiple SNPs in a candidate gene. *Genetic epidemiology*, 31(5):383–395, 2007.

[4] JP Imhof. Computing the distribution of quadratic forms in normal variables. *Biometrika*, pages 419–426, 1961.

[5] S Lagakos. Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. *Statistics in medicine*, 7(1-2):257–274, 1988.

[6] Miao-Xin Li, Hong-Sheng Gui, Johnny Kwan, and Pak Sham. GATES: a rapid and powerful gene-based association test using extended simes procedure. *American journal of human genetics*, 88(3):283–293, 2011.

[7] Mingyao Li, Kai Wang, Struan Grant, Hakon Hakonarson, and Chun Li. ATOM: a powerful gene-based association test by combining optimally weighted markers. *Bioinformatics (Oxford, England)*, 25(4):497–503, 2009.

[8] J.Z. Liu, A.F. Mcrae, D.R. Nyholt, S.E. Medland, N.R. Wray, K.M. Brown, N.K. Hayward, G.W. Montgomery, P.M. Visscher, N.G. Martin, et al. A versatile gene-based test for genome-wide association studies. *The American Journal of Human Genetics*, 87(1):139-145, 2010.

[9] E. Mangold, K. U Ludwig, S. Birnbaum, C. Baluardo, M. Ferrian, S. Herms, H. Reutter, N. A de Assis, T. Al Chawa, M. Mattheisen, et al. Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nature genetics*, 42(1):24-26, 2009.

[10] V. Moskvina, K.M. Schmidt, A. Vedernikov, M.J. Owen, N. Craddock, P. Holmans, and M.C. O'Donovan. Permutation-based approaches do not adequately allow for linkage disequilibrium in gene-wide multi-locus association analysis. *European Journal of Human Genetics*, 20, 2012.

[11] Sreekumar G Pillai, Dongliang Ge, Guohua Zhu, Xiangyang Kong, Kevin V Shianna, Anna C Need, Sheng Feng, Craig P Hersh, Per Bakke, Amund Gulsvik, et al. A genome-wide association study in chronic obstructive pulmonary disease (copd): identification of two major susceptibility loci. *PLoS genetics*, 5(3):e1000421, 2009.

[12] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W. De Bakker, M.J. Daly, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559-575, 2007.

[13] L.A. Stefanski and R.J. Carroll. Score tests in generalized linear measurement error models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 345–359, 1990.

[14] Kai Wang and Diana Abbott. A principal components regression approach to multilocus genetic association studies. *Genetic epidemiology*, 32(2):108–118, 2008.

[15] Tao Wang and Robert Elston. Improved power by use of a weighted score test for linkage disequilibrium mapping. *American journal of human genetics*, 80(2):353–360, 2007.

Testing for odds ratio bias in case-control studies

David M. Swanson and Rebecca Betensky

Abstract

Survival bias is a long-recognized problem in case-control studies, and many varieties of bias can come under this umbrella term. We focus on one of them, termed Neyman's bias or "prevalence-incidence bias." It occurs in case-control studies when exposure affects both disease and disease-induced mortality, and we give a formula for the observed, biased odds ratio under such conditions. We compare our result with previous investigations into this phenomenon and consider models under which this bias may or may not be significant. Finally, we propose three hypothesis tests to identify when Neyman's bias may be present in case-control studies. We apply these tests to two data sets, one of stroke mortality and another of brain cancer, and find some evidence of Neyman's bias in both cases.

1 Introduction

Survival bias is a frequent source of concern in case-control studies [16, 15]. Sackett describes nine types of bias common in case-control studies, and we focus our investigation on one of them, first identified by Jerzy Neyman and now known as Neyman's bias or "prevalence-incidence bias" [13]. It is a bias that occurs when exposure affects disease and disease-associated mortality and prevalent cases are sampled. Since Neyman's article was written in the 1950's when the relationship between smoking and lung cancer was under debate, he uses an example that focuses on that subject. He disregards competing risks and supposes that if, in fact, smoking is protective against lung cancer, but lung cancer mortality is far higher among non-smokers than smokers, then the odds ratio would suggest that smoking is a risk factor for disease as was being observed at the time.

Despite Neyman's early identification of this bias, methodological investigation into it has been limited. Hill (2003) uses a compartment model to show how bias arises when performing case-control studies on prevalent cases if the risk factor impacts both disease and mortality from disease. He also shows that any impact of the risk factor on mortality from other causes does not impact the observed odds ratio, which demonstrates that Neyman was justified in ignoring competing risks.

Anderson et al. (2011) performed a computational investigation into Neyman's bias, recognizing that genome-wide association studies (GWAS) and their use of prevalent cases in case-control study designs were susceptible to it. If an allele was a risk factor for both disease and mortality from disease, then the common practice of calculating an odds ratio from prevalent cases and controls could lead to biased inference. Since the odds ratios in such studies are

usually small, differences in disease-associated mortality between the exposed and unexposed would not be required for a risk allele to be observed as protective, or vice versa. Their own investigation was motivated by a locus found to be significantly associated with ischemic stroke in longitudinal studies that did not replicate using a case-control design. As a solution, they simulated data under different disease and mortality risk models and then fit regression models for % bias of the odds ratio to the disease and mortality risk model parameters. These fitted models gave researchers a means to investigate the potential biases of estimated odds ratios in their own studies.

In an editorial response to Andersen et al.'s work, [7] noted that the bias under investigation is more than anything an issue of study design. True case-control studies should be ones focused on incident, not prevalent, cases, and when lacking knowledge of the time of disease onset, one can never assume that a measure of association is unbiased. So while the work of [1] is valid, emphasis should be placed on conducting a well-executed case-control study in the first place, not attempting to assess and correct bias of a poorly conducted one. In this paper, we examine the issue of Neyman's bias from a modeling perspective and suggest methods to assess whether Neyman's bias is present in a study.

2 Methods

2.1 Background

Assume that we have a setting similar to that described in [1], where we have some binary risk SNP or gene, G, that takes on the value 1 with probability p ("exposed") and 0 with probability $1-p$ ("unexposed").

Suppose G has some unspecified association with $M_{a,i}$, $i = 1, \ldots, n$, denoting age at mortality from all other causes not associated with disease. That is, by definition of $M_{a,i}$, $i = 1, \ldots, n$, $M_{a,i} \perp\!\!\!\perp (X \quad D)^T \mid G$, where $W \perp\!\!\!\perp Y \mid Z$ denotes statistical independence of W and Y conditional on Z, D is age at disease-onset, and X is time from disease to the first disease-associated mortality cause. We describe X in greater detail later. Define $M_a \equiv \min\{M_{a,i}\}$ so that we have $M_a \perp\!\!\!\perp (X \quad D)^T \mid G$.

Additionally, suppose that G may be associated with D, again age at disease, and $M_{d,i} \equiv D + X_i$, where X_i, $i = 1, \ldots, m$, is time to the i^{th} mortality cause from disease onset and may or may not be associated with D

So we have that $M_{d,i}$ is age at the i^{th} disease-associated mortality and, depending on the joint distribution $(D \;\; X_i)^T$, $M_{d,i}$ may or may not be associated with D. If $X_i \perp\!\!\!\perp D$, then $M_{d,i}$ is necessarily associated with D because $M_{d,i} \equiv D + X_i$. In fact, we need X_i associated with D in a specific way to have $M_{d,i} \perp\!\!\!\perp D$. We do not assume X_i is a positive random variable so that we can have $P(M_{d,i} < D) \geq 0$. While it may seem counterintuitive to allow for disease-associated mortality prior to disease, this flexibility fits into a realistic framework. For example, if the disease of interest is stroke, and there exists an association between death from myocardial infarction and stroke, then indeed mortality associated with disease, though not directly caused by it, can occur before disease and can bias

the odds ratio as is shown later. Now let us define $X \equiv \min\{X_i\}$, and similarly, $M_d \equiv D + X = D + \min\{X_i\}$.

2.2 Formulae

Suppose we perform a case-control study of prevalent cases at age t^*, and define $C_a \equiv I(t^* \leq M_a)$, $C_d \equiv I(t^* \leq M_d)$, and $C \equiv \min\{C_d, C_a\}$, $I(\cdot)$ the indicator function, so that a subject can be thought of as observed if $C = 1$; i.e., the subject's event time occurs before censoring by both all-cause mortality and disease-associated mortality. Denote the cumulative distribution function associated with random variable X as $F_X(t)$. Then the target odds ratio among the population at age t^* is

$$\begin{aligned} OR_{tr}(t^*) &= \frac{pr(\text{Case, Exposed})\, pr(\text{Control, Unexposed})}{pr(\text{Control, Exposed})\, pr(\text{Case, Unexposed})} \\ &= \frac{pr(D \leq t^*, G = 1) pr(D > t^*, G = 0)}{pr(D > t^*, G = 1) pr(D \leq t^*, G = 0)} \\ &= \frac{F_{D|G=1}(t^*)\, p\, \{1 - F_{D|G=0}(t^*)\}\, (1-p)}{\{1 - F_{D|G=1}(t^*)\}\, p\, F_{D|G=0}(t^*)\, (1-p)} \\ &= \frac{F_{D|G=1}(t^*)\, \{1 - F_{D|G=0}(t^*)\}}{\{1 - F_{D|G=1}(t^*)\}\, F_{D|G=0}(t^*)}. \end{aligned}$$

Whereas, putting no constraints on the joint model $(X \quad D \quad G)^T$, the observed odds ratio among prevalent cases at age t^* is

$$\begin{aligned} OR_{ob}(t^*) &= \frac{pr(\text{Case, Exposed, Observed})\, pr(\text{Control, Unexposed, Observed})}{pr(\text{Control, Exposed, Observed})\, pr(\text{Case, Unexposed, Observed})} \\ &= \frac{pr(D \leq t^*, G = 1, C = 1)\, pr(D > t^*, G = 0, C = 1)}{pr(D > t^*, G = 1, C = 1)\, pr(D \leq t^*, G = 0, C = 1)} \\ &= \frac{pr(D \leq t^*, G = 1, C_a = 1, C_d = 1)\, pr(D > t^*, G = 0, C_a = 1, C_d = 1)}{pr(D > t^*, G = 1, C_a = 1, C_d = 1)\, pr(D \leq t^*, G = 0, C_a = 1, C_d = 1)}. \end{aligned}$$

Consider the term $pr(D \leq t^*, G = 1, C_a = 1, C_d = 1)$.

We can factor the probability as

$$\begin{aligned} & pr(D \leq t^*, G = 1, C_a = 1, C_d = 1) \\ =& pr(D \leq t^*, C_d = 1 | C_a = 1, G = 1)\, pr(C_a = 1 | G = 1)\, pr(G = 1). \end{aligned}$$

Since $M_a \perp\!\!\!\perp (X \quad D)^T \mid G$ and $M_d \equiv X + D$, $M_a \perp\!\!\!\perp M_d \mid G$, and since C_a and C_d are functions of only M_a and

M_d (with fixed and known t^*), respectively, $(D \;\; C_d) \perp\!\!\!\perp C_a \mid G$. Using this conditional independence, we have

$$\begin{aligned}
&pr(D \le t^*, C_d = 1 | C_a = 1, G = 1)\, pr(C_a = 1 | G = 1)\, pr(G = 1) \\
=&pr(D \le t^*, C_d = 1 | G = 1)\, pr(C_a = 1 | G = 1)\, pr(G = 1) \\
=&\int_0^{t^*} \{1 - F_{X|D=t,G=1}(t^* - t)\} \partial F_{D|G=1}(t) \{1 - F_{M_a|G=1}(t^*)\}\, p.
\end{aligned}$$

Reducing the other terms of $OR_{ob}(t^*)$ in the corresponding way, we have

$$= \frac{\int_0^{t^*} \{1 - F_{X|D=t,G=1}(t^* - t)\} \partial F_{D|G=1}(t) \{1 - F_{M_a|G=1}(t^*)\}\, p}{\int_{t^*}^{\infty} \{1 - F_{X|D=t,G=1}(t^* - t)\} \partial F_{D|G=1}(t) \{1 - F_{M_a|G=1}(t^*)\}\, p} \frac{\int_{t^*}^{\infty} \{1 - F_{X|D=t,G=0}(t^* - t)\} \partial F_{D|G=0}(t) \{1 - F_{M_a|G=0}(t^*)\}\, (1-p)}{\int_0^{t^*} \{1 - F_{X|D=t,G=0}(t^* - t)\} \partial F_{D|G=0}(t) \{1 - F_{M_a|G=0}(t^*)\}\, (1-p)}$$

$$= \frac{\left[\int_0^{t^*} \{1 - F_{X|D=t,G=1}(t^* - t)\} \partial F_{D|G=1}(t)\right] \left[\int_{t^*}^{\infty} \{1 - F_{X|D=t,G=0}(t^* - t)\} \partial F_{D|G=0}(t)\right]}{\left[\int_{t^*}^{\infty} \{1 - F_{X|D=t,G=1}(t^* - t)\} \partial F_{D|G=1}(t)\right] \left[\int_0^{t^*} \{1 - F_{X|D=t,G=0}(t^* - t)\} \partial F_{D|G=0}(t)\right]}. \tag{1}$$

Consider $X \perp\!\!\!\perp D \mid G$, the case when time from disease to the first disease-associated mortality, X, is independent of age at disease, D, conditional on exposure, G. The assumption may be reasonable for some exposures that are risk factors for diseases whose course is independent of the age of onset given G. In this case, we observe

$$\begin{aligned}
OR_{ob}(t^*) &= \frac{pr(\text{Case, Exposed, Observed})\, pr(\text{Control, Unexposed, Observed})}{pr(\text{Control, Exposed, Observed})\, pr(\text{Case, Unexposed, Observed})} \\
&= \frac{\left[\int_0^{t^*} \{1 - F_{X|G=1}(t^* - t)\} \partial F_{D|G=1}(t)\right] \left[\int_{t^*}^{\infty} \{1 - F_{X|G=0}(t^* - t)\} \partial F_{D|G=0}(t)\right]}{\left[\int_{t^*}^{\infty} \{1 - F_{X|G=1}(t^* - t)\} \partial F_{D|G=1}(t)\right] \left[\int_0^{t^*} \{1 - F_{X|G=0}(t^* - t)\} \partial F_{D|G=0}(t)\right]}.
\end{aligned}$$

Return to (1), and let us consider ways in which $OR_{ob}(t^*) = OR_{tr}(t^*)$ holds. Recall that $M_d \equiv D + X$, where X need not be a positive random variable. Suppose that $X \equiv A - D$, for some positive random variable A independent of D, conditional on G. Then $M_d \equiv D + X = D + (A - D) = A$. So $M_d = A$ and is independent of D given G, or in notation, $M_d \perp\!\!\!\perp D \mid G$. Notice that when M_d is defined in this way, an association necessarily exists between X and D, conditional on G, since X is itself a function of D. If $M_d \perp\!\!\!\perp D \mid G$ holds, then (1) reduces to

$$\begin{aligned}
OR_{ob}(t^*) =& \frac{\left[\{1 - F_{M_d|G=1}(t^*)\} \int_0^{t^*} \partial F_{D|G=1}(t)\right] \left[\{1 - F_{M_d|G=0}(t^*)\} \int_{t^*}^{\infty} \partial F_{D|G=0}(t)\right]}{\left[\{1 - F_{M_d|G=1}(t^*)\} \int_{t^*}^{\infty} \partial F_{D|G=1}(t)\right] \left[\{1 - F_{M_d|G=0}(t^*)\} \int_0^{t^*} \partial F_{D|G=0}(t)\right]} \\
=& \frac{F_{D|G=1}(t^*)\, 1 - F_{D|G=0}(t^*)}{1 - F_{D|G=1}(t^*)\, F_{D|G=0}(t^*)} = OR_{tr}(t^*),
\end{aligned} \tag{2}$$

where (2) uses $F_{X|D=t,G=g}(t^*-t) = F_{X+t|D=t,G=g}(t^*) = F_{X+D|D=t,G=g}(t^*) = F_{M_d|D=t,G=g}(t^*) = F_{M_d|G=g}(t^*)$. So when $M_d \perp\!\!\!\perp D \mid G$, M_d functions as M_a in the sense that $OR_{ob}(t^*)$ is no longer a function of the distribution of M_d and $OR_{tr}(t^*) = OR_{ob}(t^*)$. While $M_d \perp\!\!\!\perp D \mid G$ is a sufficient condition for $OR_{tr}(t^*) = OR_{ob}(t^*)$, it is not necessary; there exist multivariate distributions $(X \;\; D \;\; G)^T$ such that $OR_{tr}(t^*) = OR_{ob}(t^*)$, but $M_d \perp\!\!\!\perp D \mid G$ does not hold. In the hypothesis testing section of this paper, proposed tests attempt to detect deviations from $M_d \perp\!\!\!\perp D \mid G$.

3 Observations

We make a few observations based on the observed and target odds ratios. We distinguish between what we will term the "scientific null hypothesis," that at some time t^*, $OR_{tr}(t^*) = 1$, and the "bias null hypothesis," that at some time t^*, $OR_{tr}(t^*) = OR_{ob}(t^*)$. The alternative hypothesis in both cases is the complement of the null hypothesis. The observations are categorized in terms of whether the scientific null or alternative hypothesis is assumed.

First, under both the scientific null (i.e., $OR_{tr}(t^*) = 1$) and scientific alternative hypothesis (i.e., $OR_{tr}(t^*) \neq 1$), even if mortality from other causes, M_a, depends on G, it does not affect the bias of the observed odds ratio; in other words, $OR_{ob}(t^*)$ and $OR_{tr}(t^*)$ are not a function of the distribution of M_a. Thus, we may assume, as Neyman does in his original example and Hill (2003) confirms, that mortality from other causes is not present and death can only occur from disease. Similarly, we see that p, probability of exposure, does not affect $OR_{ob}(t^*)$. Also, and as expected, if $F_{X|G=g}(t^*) = 0$ for $g \in \{0, 1\}$ (which is the case when no disease-associated mortality occurs prior to t^*), then $OR_{ob}(t^*)$ is unbiased: $OR_{ob}(t^*) = OR_{tr}(t^*)$. This result is expected since it is disease-related mortality that results in the bias-inducing differential selection between the exposed and unexposed.

Secondly, and under only the scientific alternative hypothesis (i.e., $OR_{tr}(t^*) \neq 1$), if we suppose the following conditions $1 - 4$, then bias exists (i.e., $OR_{ob}(t^*) \neq OR_{tr}(t^*)$):

1. $F_{X|D,G=0}(t^* - t) = F_{X|G=0}(t^* - t) = F_{X|G=1}(t^* - t) = F_{X|D,G=1}(t^* - t)$ for all t (i.e., the mortality distribution from disease-onset is identical between the exposed and unexposed and not dependent on age at disease-onset).

2. $F_{X|G=g}(t^{**}) > 0$ for some $g \in \{0, 1\}$ (i.e., either the exposed or unexposed have positive probability of dying from disease by t^{**}, where t^{**} is defined as the time between t^* and the first possible presence of disease among the exposed or unexposed so that the bias-inducing event will have some chance of occurring prior to study at t^*).

3. $pr(X > 0) = 1$, implying $pr(D < M_d) = 1$.

4. $F_{D|G=0}(x) = F_{D|G=1}(x - k)$ for all x for some $k \neq 0$, and $F_{D|G=0}(t^*) > 0$ or $F_{D|G=1}(t^*) > 0$ (i.e., the disease distributions for the exposed and unexposed are in the same location family, and $k \neq 0$ implies

$OR_{tr}(t^*) \neq 1$).

These assumptions seem plausible if some exposure translates the mean age of disease, though the shape of the disease distribution is approximately the same between exposed and unexposed, and after disease occurrence, hazard of mortality is identical among those with and without the exposure and not a function of age at disease onset. The theorem and proof of this result is found in the *Proof and examples* section (Theorem 1). Additionally, in that proof we find that when $OR_{tr}(t^*) < 1$, then $OR_{ob}(t^*) > OR_{tr}(t^*)$, and when $OR_{tr}(t^*) > 1$, then $OR_{ob}(t^*) < OR_{tr}(t^*)$. Thus, if the degree of bias is relatively small, then it can be viewed as a bias toward an observed odds ratio of 1. However, $OR_{ob}(t^*)$ is by no means bounded by 1 and so if the amount of bias is great, $OR_{ob}(t^*)$ and $OR_{tr}(t^*)$ can lie on opposite sides of 1, leading to wrongly inferring a truly protective exposure as a risk factor for the outcome or a true risk factor as protective against the outcome.

This result of $OR_{ob}(t^*) \neq OR_{tr}(t^*)$ will not necessarily hold if conditions $1-3$ hold, but condition 4 is not satisfied (the distributions of disease of exposed and unexposed are not in the same location family). Under such a scenario, there may not be bias as Example 1 in the *Proof and examples* section illustrates. Additionally, if we only assume that conditions $2-3$ are satisfied, then there may or may not be bias. See Examples 2 and 3 in the *Proof and examples* section for instances of $OR_{ob}(t^*) = OR_{tr}(t^*)$ and $OR_{ob}(t^*) \neq OR_{tr}(t^*)$, respectively, when X is associated with G (but is independent of D given G: $X \perp\!\!\!\perp D \mid G$). It follows that if there exist no conditional independences, one can make no conclusions regarding the relationship between $OR_{tr}(t^*)$ and $OR_{ob}(t^*)$ as there is even greater flexibility in the joint model. Lastly, if only $X \perp\!\!\!\perp G \mid D$ is assumed so that X may depend on D (i.e., time to disease-induced mortality may depend on age at disease-onset), again $OR_{tr}(t^*)$ and $OR_{ob}(t^*)$ may or may not be equal. This result follows from the proof with location families and Example 1 because they are special cases of only assuming $X \perp\!\!\!\perp G \mid D$.

Third and lastly, we compare $OR_{ob}(t^*)$ and $OR_{tr}(t^*)$ under the scientific null hypothesis, $OR_{tr}(t^*) = 1$. If we only assume that $OR_{tr}(t^*) = 1$ with no conditions on $OR_{tr}(t)$ for $t < t^*$, and also that $X \perp\!\!\!\perp D \mid G$ and $F_{X|G=0}(t) \neq F_{X|G=1}(t)$ for some $t < t^{**}$, one cannot conclude anything regarding the relationship between $OR_{tr}(t^*)$ and $OR_{ob}(t^*)$. Consider Examples 4 and 5 in the *Proof and examples* section for instances of $OR_{ob}(t^*) = OR_{tr}(t^*) = 1$ and $OR_{ob}(t^*) \neq OR_{tr}(t^*) = 1$, respectively. We also observe that if $OR_{tr}(t) = 1$ for all $t \leq t^*$ and $F_{X|D,G=0}(t) = F_{X|D,G=1}(t)$ for all $t < t^{**}$, $OR_{tr}(t^*) = OR_{ob}(t^*) = 1$.

4 The odds ratio when T^* is not fixed

If the case-control study consists of people of many ages, then t^*, previously considered fixed, can be considered random. Let us denote this random variable T^*. Under these conditions, the target odds ratio becomes

$$\begin{aligned} OR_{tr}(T^*) &= \frac{pr(\text{Case, Exposed})\, pr(\text{Control, Unexposed})}{pr(\text{Control, Exposed})\, pr(\text{Case, Unexposed})} \\ &= \frac{pr(D \leq T^*, G=1)\, pr(D > T^*, G=0)}{pr(D > T^*, G=1)\, pr(D \leq T^*, G=0)} \\ &= \frac{\int F_{D|G=1}(u)\, \partial F_{T^*}(u)\, p \left\{1 - \int F_{D|G=0}(u)\, \partial F_{T^*}(u)\right\} (1-p)}{\left\{1 - \int F_{D|G=1}(u)\, \partial F_{T^*}(u)\right\} p \int F_{D|G=0}(u)\, \partial F_{T^*}(u)\, (1-p)} \\ &= \frac{\int F_{D|G=1}(u)\, \partial F_{T^*}(u) \left\{1 - \int F_{D|G=0}(u)\, \partial F_{T^*}(u)\right\}}{\left\{1 - \int F_{D|G=1}(u)\, \partial F_{T^*}(u)\right\} \int F_{D|G=0}(u)\, \partial F_{T^*}(u)}. \end{aligned}$$

Making no assumptions about the joint model $(D \;\; X \;\; M_a \;\; G)^T$, the observed odds ratio is

$$\begin{aligned} OR_{ob}(T^*) &= \frac{pr(\text{Case, Exposed, Observed})\, pr(\text{Control, Unexposed, Observed})}{pr(\text{Control, Exposed, Observed})\, pr(\text{Case, Unexposed, Observed})} \\ &= \frac{pr(D \leq T^*, T^* < M_d, T^* < M_a, G=1)\, pr(D > T^*, T^* < M_d, T^* < M_a, G=0)}{pr(D > T^*, T^* < M_d, T^* < M_a, G=1)\, pr(D \leq T^*, T^* < M_d, T^* < M_a, G=0)} \end{aligned}$$

$$= \frac{pr(D \leq T^*, T^* < M_d, T^* < M_a | G=1)\, pr(G=1)}{pr(D > T^*, T^* < M_d, T^* < M_a | G=1)\, pr(G=1)} \frac{pr(D > T^*, T^* < M_d, T^* < M_a | G=0)\, pr(G=0)}{pr(D \leq T^*, T^* < M_d, T^* < M_a | G=0)\, pr(G=0)}$$

$$= \frac{pr(D \leq T^*, T^* < M_d, T^* < M_a | G=1)\, pr(D > T^*, T^* < M_d, T^* < M_a | G=0)}{pr(D > T^*, T^* < M_d, T^* < M_a | G=1)\, pr(D \leq T^*, T^* < M_d, T^* < M_a | G=0)}$$

While $pr(G=1) = p$ cancels from $OR_{ob}(T^*)$ as before with $OR_{ob}(t^*)$, we see that even if $(D \;\; X) \perp\!\!\!\perp M_a \mid G$, we cannot factor $pr(T^* < M_a | G = g)$ out of the expression. So $OR_{ob}(T^*)$ becomes a function of M_a, causes of mortality unassociated with the disease under investigation. Additionally, regardless of whether $pr(T^* < M_a | G = g)$ factors out of the expression, $D \perp\!\!\!\perp M_d \mid G$, which we have stated before as being sufficient for $OR_{ob}(t^*) = OR_{tr}(t^*)$, is not sufficient for $OR_{ob}(T^*) = OR_{tr}(T^*)$. This point is relevant as we propose hypothesis tests below.

5 Testing

5.1 Description

We develop three methods for testing for the presence of Neyman's bias in a study. Again, the "bias null hypothesis" of these tests is $OR_{tr}(t^*) = OR_{ob}(t^*)$, and the alternative is $OR_{tr}(t^*) \neq OR_{ob}(t^*)$. While power may vary as a function of $OR_{tr}(t^*)$, the tests we propose are valid under all values of $OR_{tr}(t^*)$. Each of these three methods makes use of characteristics unique to the data when Neyman's bias is absent, and each test may be more fitting to use than the other two under certain study designs. So, for example, Tests 1 and 2 require study observations to have some variation in age at study entry, a random variable we denote T^*, while Test 3 does not, though Test 3 requires external knowledge of population prevalence of disease and exposure, while neither Test 1 nor Test 2 do so.

We have demonstrated above that $M_d \perp\!\!\!\perp D \mid G$ is a sufficient condition for $OR_{tr}(t^*) = OR_{ob}(t^*)$. Ideally, we would have data on all of D, M_d, and G and could test for conditional independences. However, in practice, it may be unlikely that one would have follow-up data on controls, in which case M_d would be unknown for subjects with $T^* < D$, and perhaps M_d would be unknown for controls as well. Thus, we propose these tests with real-world data limitations in mind.

The first two hypothesis tests we propose attempt to test whether this independence condition holds. Both of these hypothesis tests make use of previous work coming from the truncation methodology literature for tests of "quasi-independence," which refers to independence of random variables in a certain "observable" region of their joint distribution, which we explain further below [12, 2, 18].

The last hypothesis test we propose assumes $pr(D < M_d) = 1$, which may be unreasonable in some settings, but reasonable in others, and depends on whether causes of mortality associated with disease can come before disease onset. The test uses the fact that using data collected under a case-control study design along with population disease prevalence, one can estimate the population exposure proportion. If one has knowledge of the true exposure proportion, any comparison between the true, known value and the calculated quantity can reveal bias in the odds ratio from which it was calculated. Thus, in contrast to the first two tests which detect a sufficient, though not necessary, condition of unbiasedness, this latter test has power above the type 1 error whenever bias is present.

5.2 Test 1: testing for "quasi-independence" under double truncation

Tests of association using U-statistics have been proposed for double truncation settings, whereby a realization of a multivariate random variable is only observed if left and right truncating events are satisfied [2, 12]. For example, a double truncation setting would be one where W is only in the sample if W satisfies $L < W$, the left truncating event, and $W < R$, the right truncating event. Here we modify a non-parametric association test of [2], whose null hypothesis assumes in our context mutual independence of D, T^*, M_d. However, in our setting, we test only for independence of

D and M_d given G, and our observable region is where $D < T^* < M_d$ given G; i.e., realizations of observed (because $T^* < M_d$) cases (because $D < T^*$) of a given exposure status. This is a valid approach to testing $D \perp\!\!\!\perp M_d \mid G$, which is sufficient for no Neyman's bias, because $D \perp\!\!\!\perp M_d \mid G$ necessarily implies independence in the region we are defining as observable, $D < T^* < M_d$ given G. Additionally, we focus on cases under the assumption that follow-up data on M_d is more likely to be available among them. While the power of this test may suffer in comparison to one that makes use of all observations, the approach makes fewer assumptions on data availability, and in settings where $P(D < M_d)$ is close to 1, power will not suffer significantly.

To implement the hypothesis test, first we categorize all causes of mortality as M_d since if D and M_d are associated given G, and $D \perp\!\!\!\perp M_a \mid G$, then categorizing M_a as M_d will maintain that association and avoid the need to censor observations. Also, if $D \perp\!\!\!\perp M_d \mid G$ and $D \perp\!\!\!\perp M_a \mid G$, categorizing M_a as M_d will maintain $D \perp\!\!\!\perp M_d \mid G$. This approach is also legitimate from the perspective that M_d was originally defined as causes of mortality potentially, though not necessarily, associated with disease. Now suppose that we have $1, \ldots, n$ realizations of $(G_i \; D_i \; T_i^* \; M_{d,i})^T$, and that $C_{ij}^0 = 1$ (alternatively, $C_{ij}^1 = 1$) if $G = 0$ (alternatively, $G = 1$) and $\max\{D_i, D_j\} \leq \min\{T_i^*, T_j^*\}, \max\{T_i^*, T_j^*\} \leq \min\{M_{d,i}, M_{d,j}\}$, the comparability criterion, is satisfied, and $C_{ij}^0 = 0$ (alternatively, $C_{ij}^1 = 0$) otherwise. Define $n_0 \equiv \sum_{i=1}^{n-1} \sum_{j=(i+1)}^{n} C_{ij}^0$ and $n_1 \equiv \sum_{i=1}^{n-1} \sum_{j=(i+1)}^{n} C_{ij}^1$.
The test statistic for the unexposed ($G = 0$), T_0, is

$$T_0 = \frac{1}{n_0} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{sgn}\{(D_i - D_j)(M_{d,i} - M_{d,j})\} \, C_{ij}^0,$$

while the test statistic for the exposed ($G = 1$), T_1, is

$$T_1 = \frac{1}{n_1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{sgn}\{(D_i - D_j)(M_{d,i} - M_{d,j})\} \, C_{ij}^1.$$

Then $T_0 \sim N(0, v_0)$ and $T_1 \sim N(0, v_1)$, where

$$v_g = E(\text{sgn}\{(D_1 - D_2)(M_{d,1} - M_{d,2}) \, (D_1 - D_3)(M_{d,1} - M_{d,3})\} \, C_{12}^g \, C_{13}^g \mid G = g) - (\tau_{D,M_d}^g \mu_{D,M_d})^2$$

with $g \in \{0, 1\}$, and where $\tau_{D,M_d}^g = E(\text{sgn}\{(D_1 - D_2)(M_{d,1} - M_{d,2})\} \mid C_{12} = 1, G = g)$ and $\mu_{D,M_d} = pr(C_{12} = 1)$, with $\text{sgn}(x) = 1$ for $x > 0$, -1 for $x < 0$, and 0 for $x = 0$.

Since we would reject if either T_0 or T_1 falls in some predetermined critical region because dependence between D and M_d given either $G = 0$ or $G = 1$ may mean $OR_{tr}(t^*) \neq OR_{ob}(t^*)$, in order to achieve a size α test, we can use a p-value threshold of α^* for T_0 and T_1, where α^* satisfies the equation $\alpha = 1 - (1 - \alpha^*)^2$. So we propose a test that rejects for $\max\{\text{abs}(T_0/v_0^{1/2}), \text{abs}(T_1/v_1^{1/2})\} > z_{1-\alpha^*/2}$, where $\text{abs}(x)$ denotes the absolute value of x and

$z_{1-\alpha^*/2}$ is the $(z_{1-\alpha^*/2})^{th}$ quantile of a standard normal random variable.

Also, since $D \perp\!\!\!\perp M_d \mid G$ is a subset of situations for which $OR_{tr}(t^*) = OR_{ob}(t^*)$, our test is likely conservative, though not overly so assuming that the majority of situations in which $OR_{tr}(t^*) = OR_{ob}(t^*)$ holds result from $D \perp\!\!\!\perp M_d \mid G$ being satisfied. Power curves for Test 1 as a function of the association between D and M_d are shown in Figures 12 and 13.

It is important here to make a reference to our observations regarding $OR_{ob}(T^*)$ and $OR_{tr}(T^*)$, the odds ratio when T^* is viewed as random. In Section (4) above, we saw that the distribution for M_a did not factor out of the odds ratio even when $(D \;\; X) \perp\!\!\!\perp M_a \mid G$, and that additionally even under the assumption of $D \perp\!\!\!\perp M_d \mid G$, whether or not the previous assumption held, $OR_{ob}(T^*)$ could be biased; we needed a fixed t^* for these conditional independencies to result in $OR_{tr}(t^*) = OR_{ob}(t^*)$. Thus, it may seem illogical to be proposing a test that requires variation in T^*, which is precisely when the odds ratio will almost certainly be biased as shown in Section (4). If we do find that $D \perp\!\!\!\perp M_d \mid G$, sufficient for no Neyman's bias, we would need to then stratify our sample according to similar values of T^* such that, within each stratum, T^* can be effectively considered fixed and then calculate the odds ratio for these different strata. We could then combine these strata into a average odds ratio if desirable or just consider each stratum-specific odds ratio separately. This observation is also true with Test 2, where if we found $D \perp\!\!\!\perp T^* \mid G$, which we will see implies $D \perp\!\!\!\perp M_d \mid G$ under certain assumptions, then we would need to stratify by T^* in the data and calculate $T^* = t^*$ specific odds ratios to be unbiased for $OR_{tr}(t^*)$.

5.3 Test 2: testing "quasi-independence" under left truncation

We now describe a test related to Test 1, but one which does not require knowledge of M_d. Such a test might be fitting if a data set did not have follow-up on subjects, but did record age at onset of disease for cases. Groundwork for the test is based on causal directed acyclic graphs (DAGs), borrowed from the causal inference literature [8].

By nature of any observation being in the study, the event $I(T^* < M_d) = 1$ must be satisfied and is a conditioning event. Additionally, by definition of $I(T^* < M_d)$, there exists an association between it and both T^* and M_d. Thus, we see in Figures 10 and 11 arrows between these random variables, indicative of a possible association, and a square around $I(T^* < M_d)$, indicative of a conditioning event. Applying rules of d-separation, also borrowed from the causal inference literature, and assuming $0 < P(T^* < M_d) < 1$ so the conditioning event is non-trivial, we see that when there exists an arrow (i.e., a possible association) between D and M_d, D and T^* are not d-separated (or are associated, in this case), whereas when there does not exist an arrow, D and T^* are d-separated (or are not associated) [8]. So then a lack of association between D and T^* given G implies $D \perp\!\!\!\perp M_d \mid G$ according to Figures 10 and 11, which we know by previous work implies that Neyman's bias is not present. Similarly, an association between D and T^* implies that an association exists between D and M_d, which we have shown previously may mean that bias is present, except in special cases.

We could assume D and T^* are known for all observations in our data set and propose performing a test of association for these random variables under the framework described above. However, doing so is unrealistic as it assumes follow-up data on age at disease, D, for those observed at T^* as controls (i.e., those with $D > T^*$). Thus, we assume D and T^* are only observed for cases (i.e., those realizations satisfying $D < T^*$) and propose a test of "quasi-independence" between D and T^* given G in the region of $D < T^*$ given G. If we assume that the independence which holds on the region $D < T^*$ also holds for the entire joint distribution of $(D\ T^*)^T$, then since there is an association between D and T^* given G if and only if D and M_d are associated given G, this test is valid.

We describe here this proposed test of quasi-independence under only left truncation, in contrast to the double truncation setting described in Test 1 [17, 12]. As before, let there be n realizations of $(G_i\ D_i\ T_i^*)^T$, and again define B_{ij}^0 (alternatively, B_{ij}^1) similarly to how we did with C_i^0 (alternatively, C_i^1), where $B_{ij}^0 = 1$ if $G = 0$ and $\max\{D_i, D_j\} \leq \min\{T_i^*, T_j^*\}$ and $B_{ij}^0 = 0$ otherwise, and where $B_{ij}^1 = 1$ if $G = 1$ and $\max\{D_i, D_j\} \leq \min\{T_i^*, T_j^*\}$ and $B_{ij}^1 = 0$ otherwise. Also, define $m_0 \equiv \sum_{i=1}^{n-1} \sum_{j=(i+1)}^{n} B_{ij}^0$ and $m_1 \equiv \sum_{i=1}^{n-1} \sum_{j=(i+1)}^{n} B_{ij}^1$.

Then the test statistic for the unexposed ($G = 0$), W_0, is

$$W_0 = \frac{1}{m_0} \sum_{i=1}^{n-1} \sum_{j=(i+1)}^{n} \text{sgn}\{(D_i - D_j)(T_i^* - T_j^*)\}\, B_{ij}^0,$$

while the test statistic for the exposed ($G = 1$), W_1, is

$$W_1 = \frac{1}{m_1} \sum_{i=1}^{n-1} \sum_{j=(i+1)}^{n} \text{sgn}\{(D_i - D_j)(T_i^* - T_j^*)\}\, B_{ij}^1.$$

Then $W_0 \sim N(0, u_0)$ and $W_1 \sim N(0, u_1)$, where

$$u_g = E(\text{sgn}\{(D_1 - D_2)(T_1^* - T_2^*)\,(D_1 - D_3)(T_1^* - T_3^*)\}B_{12}^g \cdot B_{13}^g \,|\, G = g) - (\tau_{D,T^*}^g \mu_{D,T^*})^2$$

with $g \in \{0, 1\}$, and where $\tau_{D,T^*}^g = E(\text{sgn}\{(D_1 - D_2)(T_1^* - T_2^*)\} \mid B_{12} = 1, G = g)$ and $\mu_{D,T^*} = pr(B_{12} = 1)$. As with Test 1, since we would reject if either W_0 or W_1 falls in some predetermined critical region because dependence between D and M_d given either $G = 0$ or $G = 1$ may mean $OR_{tr}(t^*) \neq OR_{ob}(t^*)$, for a size α test, our p-value threshold α^* for W_0 and W_1 satisfies $\alpha = 1 - (1 - \alpha^*)^2$. Thus, our test rejects for $\max\{\text{abs}(W_0/u_0^{1/2}), \text{abs}(W_1/u_1^{1/2})\} > z_{1-\alpha^*/2}$. Power curves for Test 2 as a function of the association between D and M_d are shown in Figures 12 and 13. Of these two figures, Figure 12 represents a more realistic scenario in which the study sample size is constant between the two tests, though the number of comparable pairs differs. Figure 13 is of more theoretical interest, showing the power comparison if one were to somehow be able to hold the number of comparable pairs between the two tests constant. Generally, the number of comparable pairs for Test 1 will be a

strict subset of the number of comparable pairs for Test 2 because the comparability criterion is more strict for Test 1. The larger difference in power between tests observed in Figure 13 compared to Figure 12 results from Test 1 being calculated from more comparable pairs. While Test 1 seems more powerful in simulation in both figures, Test 2 may be a better test for the bias since it is the selecting event of $T^* < M_d$ that determines if any association between D and M_d is biasing the odds ratio. Test 2 may be more indicative of that event occurring in a large proportion of observations than Test 1.

As mentioned at the end of the description of Test 1 and for reasons given there, if this test does not reject $D \perp\!\!\!\perp T^* \mid G$, implying $D \perp\!\!\!\perp M_d \mid G$, we would again need to stratify the data by T^* in order for $OR_{ob}(t^*)$ to be unbiased for $OR_{tr}(t^*)$.

5.4 Test 3: estimating population exposure proportion

With knowledge of disease prevalence, we can construct an estimate of the exposure in the general population from case-control study data that is unbiased in the absence of Neyman's bias, but biased otherwise. Thus, if the exposure proportion in the population is also known, as might be the case in GWAS where minor allele frequencies (MAFs) are oftentimes known for SNPs in different populations, we can test for the presence of Neyman's bias by examining their discrepancy. We develop one possible hypothesis test below where, again, H_0 is $OR_{tr}(t^*) = OR_{ob}(t^*)$, and H_a is the complement of H_0.

If we make an assumption of $pr(D < M_d) = 1$, then in comparing $OR_{tr}(t^*)$ and $OR_{ob}(t^*)$, we see that their equivalence depends on

$$\frac{F_{D|G=1}(t^*)\left\{1 - F_{D|G=0}(t^*)\right\}}{\left\{1 - F_{D|G=1}(t^*)\right\} F_{D|G=0}(t^*)} = \frac{\left[\int_0^{t^*} \{1 - F_{X|D=t,G=1}(t^* - t)\} \partial F_{D|G=1}(t)\right] \left[\{1 - F_{D|G=0}(t^*)\}\right]}{\left[\{1 - F_{D|G=1}(t^*)\}\right] \left[\int_0^{t^*} \{1 - F_{X|D=t,G=0}(t^* - t)\} \partial F_{D|G=0}(t)\right]} \tag{3}$$

if and only if

$$\frac{F_{D|G=1}(t^*)}{F_{D|G=0}(t^*)} = \frac{\int_0^{t^*} \{1 - F_{X|D=t,G=1}(t^* - t)\} \partial F_{D|G=1}(t)}{\int_0^{t^*} \{1 - F_{X|D=t,G=0}(t^* - t)\} \partial F_{D|G=0}(t)}$$

So define $p_2(t^*) \equiv pr(G = 1 | D < t^*) = pr(G = 1 \mid \text{Case at } t^*)$. Then defining

$$h(t^*) \equiv \frac{F_{D|G=1}(t^*)}{F_{D|G=0}(t^*)} = \frac{pr(G = 1 \mid \text{Case at } t^*)}{pr(G = 0 \mid \text{Case at } t^*)} = \frac{pr(G = 1 \mid \text{Case at } t^*)}{1 - pr(G = 1 \mid \text{Case at } t^*)},$$

we have $p_2(t^*) = \frac{h(t^*)}{1+h(t^*)}$, and defining

$$h^*(t^*) \equiv \frac{\int_0^{t^*} \{1 - F_{X|D=t,G=1}(t^* - t)\} \partial F_{D|G=1}(t)}{\int_0^{t^*} \{1 - F_{X|D=t,G=0}(t^* - t)\} \partial F_{D|G=0}(t)} = \frac{pr(G = 1 \mid \text{Case at } t^*, \text{Not censored from disease by } t^*)}{pr(G = 0 \mid \text{Case at } t^*, \text{Not censored from disease by } t^*)},$$

then we have $p_2^N(t^*) \equiv pr(G = 1 \mid \text{Case at } t^*, \text{Not censored from disease by } t^*) = \frac{h^*(t^*)}{1+h^*(t^*)}$. When equation 3 does not hold,

$$p_2^N(t^*) \equiv \frac{h^*(t^*)}{1 + h^*(t^*)} \neq \frac{h(t^*)}{1 + h(t^*)} \equiv p_2(t^*).$$

Thus, if bias is present so that $OR_{tr}(t^*) \neq OR_{ob}(t^*)$, then $h(t^*) \neq h^*(t^*)$, and it will follow that $p_2(t^*) \neq p_2^N(t^*)$. This idea can be leveraged in a hypothesis test if there is external knowledge of the population exposure proportion and population prevalence of disease.

By definition of $p_2^N(t^*)$, its estimator, $\widehat{p_2}^N(t^*)$, is the observed exposure proportion among cases where $E(\widehat{p_2}^N(t^*)) = p_2^N(t^*)$. Let $p_1(t^*) \equiv pr(G = 1 \mid D > t^*, M_d > t^*)$, and since $pr(D < M_d) = 1$ by assumption, $p_1(t^*) = pr(G = 1 \mid D > t^*) = pr(G = 1 \mid \text{Control at } t^*)$. Then $\widehat{p}_1(t^*)$ is the observed exposure proportion among controls, and $E(\widehat{p}_1(t^*)) = p_1(t^*)$.

We will estimate $p^N(t^*) \equiv p_1(t^*)\{1 - p^*(t^*)\} + p_2^N(t^*)\, p^*(t^*)$ with $\widehat{p}_1(t^*)\{1 - p^*(t^*)\} + \widehat{p}_2^N(t^*)\, p^*(t^*)$. Also, define $p^*(t^*) \equiv pr(\text{Case at } t^*) = pr(D < t^*)$, which implies $\{1 - p^*(t^*)\} = pr(\text{Control at } t^*) = pr(D > t^*)$. So $p^*(t^*)$ is population prevalence of disease at a common age t^* and is considered fixed and known. Since

$$\begin{aligned} pr(G = 1) &= pr(G = 1 \mid D > t^*)\, pr(D > t^*) + pr(G = 1 | D < t^*)\, pr(D < t^*) \\ &= p_1(t^*)\{1 - p^*(t^*)\} + p_2(t^*)\, p^*(t^*), \end{aligned}$$

if $p_2(t^*) = p_2^N(t^*)$, which indicates that $OR_{ob}(t^*) = OR_{tr}(t^*)$, then $p^N(t^*) = pr(G = 1)$. Since we consider $pr(G = 1)$ fixed and known, the discrepancy between $\widehat{p}^N(t^*)$ and $pr(G = 1)$ will inform our test.

Define $\delta(t^*) = p_2(t^*) - p_2^N(t^*)$. Then

$$\begin{aligned} p^N(t^*) + \delta(t^*)\, p^*(t^*) =& p_1(t^*)\{1 - p^*(t^*)\} + p_2^N(t^*)\, p^*(t^*) + \delta(t^*)\, p^*(t^*) \\ =& p_1(t^*)\{1 - p^*(t^*)\} + p_2^N(t^*)\, p^*(t^*) + \{p_2(t^*) - p_2^N(t^*)\}\, p^*(t^*) \\ =& p_1(t^*)\{1 - p^*(t^*)\} + p_2(t^*)\, p^*(t^*) = pr(G = 1). \end{aligned}$$

So $pr(G = 1)$ and $p^N(t^*)$ differ by $\delta(t^*)\, p^*(t^*)$. The variance associated with our estimate of the exposure proportion $\widehat{p}^N(t^*)$ is

$$v \equiv var\{\widehat{p}^N(t^*)\} = \{p^*(t^*)\}^2 \left[\frac{p_2^N(t^*)\{1 - p_2^N(t^*)\}}{n_2}\right] + \{1 - p^*(t^*)\}^2 \left[\frac{p_1(t^*)\{1 - p_1(t^*)\}}{n_1}\right],$$

where n_2 is the number of cases and n_1 the number of controls. We can estimate v with $\hat{p}_1(t^*)$ and $\widehat{p}_2^N(t^*)$ and call the quantity $\hat{v}$. So using a large sample approximation, we can construct an α level hypothesis test for the presence of Neyman's bias by rejecting for

$$\left|\frac{pr(G=1)-\widehat{p}^N(t^*)}{\hat{v}^{1/2}}\right| \sim |Z| > z_{1-\alpha/2}.$$

The power becomes

$$\left|\frac{pr(G=1)-\widehat{p}^N(t^*)}{\hat{v}^{1/2}}\right| > z_{1-\alpha/2}$$

$$\approx \frac{pr(G=1)-\{\widehat{p}^N(t^*)+\delta(t^*)p^*(t^*)\}}{\hat{v}^{1/2}} \sim Z > \{z_{1-\alpha/2}-\delta(t^*)p^*(t^*)\}/\hat{v}^{1/2}),$$

assuming one tail probability negligible. We see that power decreases as $p^*(t^*)$ decreases and increases with $\delta(t^*)$, interpreted as the "degree of Neyman's bias."

Power curves for Test 3 are shown in Figure 14. Consistent with our understanding of the teat, power increases at $p^*(t^*)$ increases. In Figure 14, we assume that there is no variation in t^* so that $p^*(t^*)$ is also fixed.

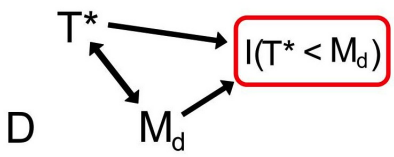

Figure 10: This DAG provides the framework for Test 2. When D is not associated with M_d, there is no association between D and T^*, despite the conditioning event, using rules of DAGs. This figure represents these random variables within each stratum of G.

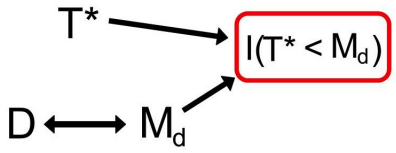

Figure 11: This DAG provides the framework for Test 2. When D is associated with M_d, an association between D and T^* is induced due to the conditioning event using rules of DAGs. This figure represents these random variables within each stratum of G.

6 Data analysis

6.1 Test 2 applied to a brain cancer data set

We apply Test 2 to a brain cancer data set. Seventy-five subjects with oligodendroglioma, a common variant of malignant brain tumors, were enrolled in a study at the London Regional Cancer Centre from 1984-1999 [3, 10]. The data set consisted of patient age at diagnosis of oligodendroglioma (i.e., age at disease, D) and age at start of chemotherapy (i.e., entry into the study, T^*) in addition to genetic markers and other covariates. We consider the marker at the 1pLOH locus, thought to potentially be associated with tumor sensitivity to chemotherapy. Applying Test 2 to the data set, first within the exposed stratum of the 1pLOH marker, we obtain a Z-statistic of 6.85, significant

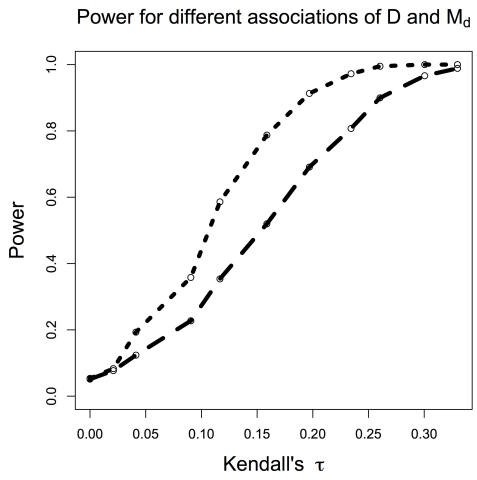

Figure 12: Comparison of power between tests 1 (short dashes) and 2 (long dashes) as a function of the association between D and M_d, measured by Kendall's τ, holding the sample size constant.

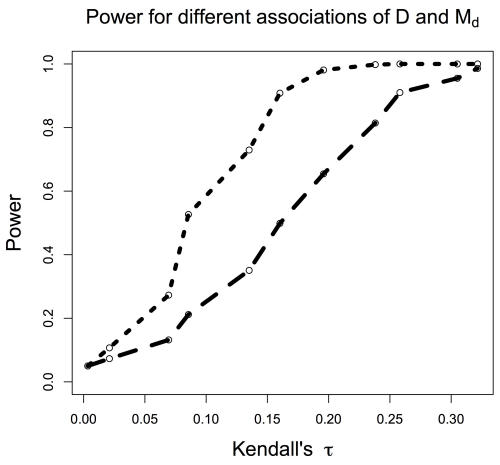

Figure 13: Comparison of power for tests 1 (short dashes) and 2 (long dashes) as a function of the D and M_d as measured by Kendall's τ, holding the number of comparable pairs constant.

at the 0.05 level ($p < 0.001$). The sample size was insufficient to apply the test to the unexposed stratum. However, since a significant test statistic within any stratum is sufficient for rejection of the null hypothesis, we reject the null hypothesis of $D \perp\!\!\!\perp M_d \mid G$ and conclude that there could be an association between D and M_d within strata of G. The result of the test suggests that if one were to calculate an odds ratio of oligodendroglioma for the 1pLOH marker at a fixed age of subjects, the result may be biased.

6.2 Test 3 applied to a stroke-mortality data set

We apply Test 3 to a GWAS data set of ischemic stroke coming from a cohort based at Massachusetts General Hospital. We use a wide interval estimate of ischemic stroke prevalence, ranging from 0.5%-5%, based on a search of the stroke literature [6, 11, 5]. With this range of $p^*(t^*)$, we reconstruct what would be population exposure proportion, which is unbiased for the true population exposure proportion assuming that Neyman's bias is not present. We calculate a test statistic based on the difference between the true population exposure proportion and our estimate of it, divided by an estimate of the standard error. Using a 0.0005 Bonferroni-adjusted significance level, we find that 42 of the 99 SNPs in the study suggest that Neyman's bias may be present. The interpretation of this result is that, were one to calculate an odds ratio for stroke mortality with any one of these 42 SNPs, that odds ratio may be biased.

7 Discussion

While our result for Test 2 with the brain cancer data suggests that Neyman's bias may be present because the within stratum association between D and T^* suggests a within stratum association of D and M_d, we should restate that an

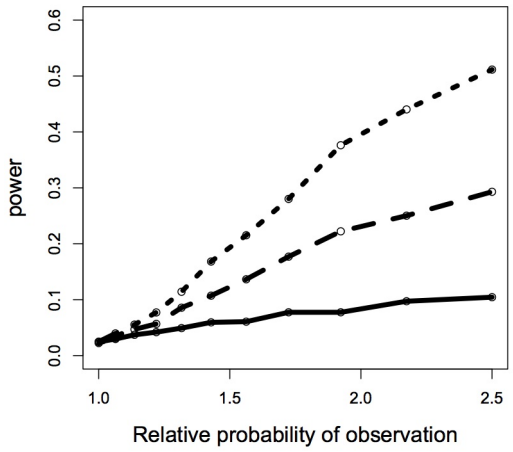

Figure 14: Power for test 3 as a function of $p^*(t^*)$ and the relative probability of observing the unexposed cases versus exposed cases. As the relative probability increases (i.e., it is more likely to observed unexposed cases than exposed cases) as is the case when there are a greater number of mortality-inducing events among the exposed, there is more bias and power. The solid, dashed, and dotted lines represent population prevalences of disease ($p^*(t^*)$) of 0.1, 0.2, and 0.3, respectively.

association within strata does not necessary imply that bias is present; it is only when the independence holds that we can conclude that Neyman's bias is not present. Additionally, the study design may contribute to a within strata association between D and T^* and so the authors suggest that more work is needed to form stronger conclusions regarding the potential presence of Neyman's bias in this study.

As with the result from Test 2, the rejection of the null hypothesis of no Neyman's bias in the stroke-mortality data by Test 3 needs confirmatory analyses. A primary concern is that if the population underlying the measurements in dbSNP, the source of our "true" population MAFs against which we compare the estimate, is significantly different than that composing the study subjects, the type 1 error could be inflated. Since for many of the SNPs in the data set, the MAF among cases and the MAF among controls did not contain the population MAF, which should be the case as the sample size gets large, there is some evidence of different underlying populations. Another assumption that may not be satisfied is $pr(D < M_d) = 1$. While $pr(D < M_d) = 1$ is unlikely to ever be fully satisfied, ischemic stroke is an event with numerous comorbidities and so violations of the assumption may be too large for a valid test [14, 4]. Lastly, description of Test 3 showed that the power for detection of bias goes to 0 as the population prevalence of disease gets small. The implication of this result is that any bias detected when population prevalence of disease ranges over a relatively small 0.5%-5% is more likely due to unsatisfied assumptions than genuine Neyman's bias.

We did not use Test 1 on the brain cancer and stroke data sets because of an insufficient sample size and insufficient covariates, respectively. The sample size was insufficient in the brain cancer data set because the comparability criterion for Test 1 is more stringent than that for Test 2 and so there are only a limited number of pairs of observations

that can contribute to estimation of the necessary parameters, especially when overlap between the multivariate random variables $(D \quad T^* \quad M_d)^T$ is minimal. Thus, while Test 2 might be thought of as somewhat removed from testing $D \perp\!\!\!\perp M_d \mid G$ because it tests $D \perp\!\!\!\perp T^* \mid G$ as a proxy for it, one advantage of Test 2 over Test 1 is that there are fewer restrictions imposed by the comparability criterion, allowing for more flexible use of the data.

8 Proof and examples

We give the proof of the direction of Neyman's bias under certain modeling assumptions and examples of when Neyman's bias does or does not occur, both referenced in the *Observations* section of this paper.

Theorem 1. *If G is associated with D such that $OR(t^*) \neq 1$, the distribution of $D \mid (G = 0)$ and $D \mid (G = 1)$ belong to the same location family, $pr(X > 0) = 1$, $pr(X < t^{**}) > 0$ (where t^{**} is defined as the time between t^* and the first possible presence of disease among the exposed or unexposed), and $X \perp\!\!\!\perp (D \quad G)^T$, then $OR_{ob}(t^*) \neq OR_{tr}(t^*)$. Specifically, if $D \mid (G = 0)$ is stochastically greater than $D \mid (G = 1)$ (alternatively, stochastically less than) so that exposure is a risk factor for disease (alternatively, protective against disease), then $OR_{ob}(t^*) < OR_{tr}(t^*)$ (alternatively, $OR_{ob}(t^*) > OR_{tr}(t^*)$).*

Proof. Define $\partial F_{D|G=0}(x)/\partial x = f_0(x)$ and $\partial F_{D|G=1}(x)/\partial x = f_1(x)$, and suppose that $f_1(x) = f_0(x - k)$ for some k positive, without loss of generality. Such a scenario corresponds to exposure being protective against disease, though below we will also consider it a risk factor. $f_1(x)$ and $f_0(x)$ are in the same location family. Define $F(x)$ as the cumulative distribution function of X evaluated at x and remember $F(0) = 0$ and $F(t^*) > 0$. Consider the two quantities:

$$\frac{\int_0^{t^*} \{1 - F(t^* - x)\} f_0(x) \partial x}{\int_0^{t^*} f_0(x) \partial x} \quad \text{and} \quad \frac{\int_0^{t^*} \{1 - F(t^* - x)\} f_1(x) \partial x}{\int_0^{t^*} f_1(x) \partial x},$$

which we call the "percent erosion" of $\int_0^{t^*} f_0(x) \partial x$ and $\int_0^{t^*} f_1(x) \partial x$, respectively. Then

$$\frac{\int_0^{t^*} \{1 - F(t^* - x)\} f_1(x) \partial x}{\int_0^{t^*} f_1(x) \partial x} = \frac{\int_0^{t^*} \{1 - F(t^* - x)\} f_0(x - k) \partial x}{\int_0^{t^*} f_0(x - k) \partial x} = \frac{\int_{-k}^{(t^* - k)} \big[1 - F\{t^* - (x + k)\}\big] f_0(x) \partial x}{\int_{-k}^{(t^* - k)} f_0(x) \partial x}.$$

Since $F(\cdot)$ a cumulative distribution function and therefore increasing, we have

$$\frac{\int_0^{t^*} \{1 - F(t^* - x)\} f_1(x) \partial x}{\int_0^{t^*} f_1(x) \partial x} = \frac{\int_{-k}^{(t^* - k)} \big[1 - F\{t^* - (x + k)\}\big] f_0(x) \partial x}{\int_{-k}^{(t^* - k)} f_0(x) \partial x} > \frac{\int_0^{t^*} \{1 - F(t^* - x)\} f_0(x) \partial x}{\int_0^{t^*} f_0(x) \partial x}, \quad (4)$$

because at every "successive" ∂x in each integral, $1 - F\{t^* - (x + k)\} \geq 1 - F(t^* - x)$ and there is some $0 < x < t^*$ for which $1 - F\{t^* - (x + k)\} > 1 - F(t^* - x)$. Thus, the "percent erosion" of $f_0(x)$ will always be greater than that of $f_1(x) = f_0(x - k)$, which is intuitive since $f_1(\cdot)$ is located to the right of $f_0(\cdot)$ and thus subject to the corrosive

effects of $F(\cdot)$ for less "time." Then using the inequality in (4),

$$\begin{aligned}
1 > &\left[\frac{\int_0^{t^*}(1-F(t^*-x))f_0(x)\partial x}{\int_0^{t^*} f_0(x)\partial x}\right] \Big/ \left[\frac{\int_0^{t^*}(1-F(t^*-x))f_1(x)\partial x}{\int_0^{t^*} f_1(x)\partial x}\right] \\
=&\frac{\int_0^{t^*} f_1(x)\ \partial x\ p}{\int_0^{t^*} f_0(x)\ \partial x\ (1-p)} \times \frac{\int_0^{t^*}\{1-F(t^*-x)\}f_0(x)\ \partial x\ (1-p)}{\int_0^{t^*}\{1-F(t^*-x)\}f_1(x)\ \partial x\ p} \\
=&\frac{pr(\text{Case, Exposed})}{pr(\text{Case, Unexposed})} \times \frac{pr(\text{Case, Unexposed, Observed})}{pr(\text{Case, Exposed, Observed})},
\end{aligned}$$

which implies that

$$\frac{pr(\text{Case, Exposed, Observed})}{pr(\text{Case, Unexposed, Observed})} > \frac{pr(\text{Case, Exposed})}{pr(\text{Case, Unexposed})} \quad \text{and} \quad OR_{ob}(t^*) > OR_{tr}(t^*)$$

since $pr(X > 0)$ implies $pr(\text{Control, Exposed, Observed}) = pr(\text{Control, Exposed})$ and $pr(\text{Control, Unexposed, Observed}) = pr(\text{Control, Unexposed})$. Again, these inequalities only hold when exposure is protective against disease. When exposure is a risk factor for disease and therefore shifts the mean age of disease onset to the left under the above assumptions,

$$\frac{pr(\text{Case, Exposed, Observed})}{pr(\text{Case, Unexposed, Observed})} < \frac{pr(\text{Case, Exposed})}{pr(\text{Case, Unexposed})} \quad \text{and} \quad OR_{ob}(t^*) < OR_{tr}(t^*)$$

using analogous results. So we see that the bias is not toward the null, but in a definite direction depending on model assumptions.

□

Example 1. Consider $D \mid (G = 1)$ uniform on $(0, 2)$, $D \mid (G = 0)$ uniform on $(0, 1)$, and X uniform on $(0, 3)$, independent of G. Clearly the distributions of disease for exposed and unexposed are not in the same location family in this case, and the model for X corresponds to disease-induced mortality necessarily occurring within 3 times units after disease, D. We need only consider cases when investigating the odds ratio since we assume $pr(X > 0) = 1$, implying $pr(D < M_d) = 1$. Taking $t^* = 1$,

$$\begin{aligned}
&\frac{pr(\text{Case, Exposed, Observed})}{pr(\text{Case, Unexposed, Observed})} = \frac{\int_0^1 (2/3 + x/3)\,(1/2)\,p\,\partial x}{\int_0^1 (2/3 + x/3)\,1\,(1-p)\,\partial x} \\
=&\frac{1/2\ \int_0^1 (2/3 + x/3)\,p\,\partial x}{1\ \int_0^1 (2/3 + x/3)\,(1-p)\,\partial x} = \frac{1\,p}{2\,(1-p)} = \frac{pr(\text{Case, Exposed})}{pr(\text{Case, Unexposed})}.
\end{aligned}$$

So we have X independent of exposure status and time of disease-onset, as was the case above, but here $OR_{ob} =$

OR_{tr}.

Example 2. Consider again $D \mid (G=1)$ uniform on $(0,2)$, and $D \mid (G=0)$ uniform on $(0,1)$. However, consider $X \mid (G=1)$ uniform on $(0,3)$ and $X \mid (G=0)$ with density $f_{X|G=0}(x) = 2/3\,(1-x)^2$ on $[0, 1+(9/2)^{1/3}]$. Again, we need only consider cases when investigating potential bias of the odds ratio since we assume $pr(D < M_d) = 1$ so that controls are not subject to the bias-inducing mortality event. Taking $t^* = 1$,

$$\frac{pr(\text{Case, Exposed, Observed})}{pr(\text{Case, Unexposed, Observed})} = \frac{\int_0^1 (2/3 + x/3)\,(1/2)\,p\,\partial x}{\int_0^1 (7/9 + 2x^3/9)\,1\,(1-p)\,\partial x}$$
$$= \frac{1/2 \cdot \int_0^1 (2/3 + x/3)\,p\,\partial x}{1\,\int_0^1 (7/9 + 2x^3/9)\,(1-p)\,\partial x} = \frac{1/2\,(5/6)\,p}{1\,(5/6)\,(1-p)} = \frac{1\,p}{2\,(1-p)} = \frac{pr(\text{Case, Exposed})}{pr(\text{Case, Unexposed})},$$

and so here we have no bias again.

Example 3. Assume the same models of D conditional on G, and suppose $X \mid (G=1)$ is uniform on $(0,3)$ and $X \mid (G=0)$ has density $f_{X|G=0}(x) = 5/2\,(1-x)^4$ on $[0, 1+2^{1/5}]$. For the reasons given above, we again only consider cases for investigating the bias of the odds ratio. Taking $t^* = 1$,

$$\frac{pr(\text{Case, Exposed, Observed})}{pr(\text{Case, Unexposed, Observed})} = \frac{\int_0^1 (2/3 + x/3)\,(1/2)\,p\,\partial x}{\int_0^1 (1/2 + x^5/2)\,1\,(1-p)\,\partial x}$$
$$= \frac{1/2\,\int_0^1 (2/3 + x/3)\,p\,\partial x}{1\,\int_0^1 (1/2 + x^5/2)\,(1-p)\,\partial x} = \frac{1/2\,(5/6)\,p}{1\,(7/12)\,(1-p)} \neq \frac{1\,p}{2\,(1-p)} = \frac{pr(\text{Case, Exposed})}{pr(\text{Case, Unexposed})},$$

and so here we have bias.

Example 4. Take $D \mid (G=1)$ with density $f_{D|G=1}(x) = x^2/4$ on $[0, 12^{1/3}]$, $D \mid (G=0)$ with density $f_{D|G=0}(x) = x/3$ $[0, 6^{1/2}]$. Then let $X \mid (G=1)$ have density $f_{X|G=1}(x) = (2-x)^2/4$ on $[0, 2+4^{1/3}]$ and $X \mid (G=0)$ be uniform on $[0,2]$. As before, we need only consider cases when investigating the odds ratio since we assume $pr(D < M_d) = 1$ so that controls are not subject to the bias-inducing mortality event. Taking $t^* = 2$,

$$\frac{pr(\text{Case, Exposed, Observed})}{pr(\text{Case, Unexposed, Observed})} = \frac{\int_0^2 (1/3 + 1/12\,x^3)\,(x^2/4)\,p\,\partial x}{\int_0^2 (x/2)\,x/3\,(1-p)\,\partial x}$$
$$= \frac{(4/9)\,p}{4/9\,(1-p)} = \frac{p\,\int_0^2 (x^2/4)\,\partial x}{(1-p)\,\int_0^2 x/3\,\partial x} = \frac{p}{1-p} = \frac{pr(\text{Case, Exposed})}{pr(\text{Case, Unexposed})}.$$

Remember that $pr(\text{Case, Exposed})/pr(\text{Case, Unexposed}) = p/(1-p)$ implies $OR_{tr}(t^*) = 1$ when $pr(D < M_d) = 1$, which is assumed from condition 3.

Example 5. On the other hand, we can obtain a biased odds ratio using the same conditional disease models as in the previous example and having $X \mid (G=1)$ with density $f_{X|G=1}(x) = (2-x)^2/4$ on $[0, 2+4^{1/3}]$ and $X \mid (G=0)$

uniform on $[0, 2]$. We again assume $pr(D < M_d) = 1$ from condition 3. Taking $t^* = 2$,

$$\frac{pr(\text{Case, Exposed, Observed})}{pr(\text{Case, Unexposed, Observed})} = \frac{\int_0^2 (1/2 + 1/16\, x^3)\, (x^2/4)\, p\, \partial x}{\int_0^2 (x/2)\, x/3\, (1-p)\, \partial x} = \frac{p\,(1/2)}{(1-p)\,4/9}$$

$$\neq \frac{(4/9)\,p}{4/9\,(1-p)} = \frac{p\,\int_0^2 (x^2/4)\,\partial x}{(1-p)\,\int_0^2 x/3\,\partial x} = \frac{p}{1-p} = \frac{pr(\text{Case, Exposed})}{pr(\text{Case, Unexposed})}.$$

9 Acknowledgement

The authors wish to thank Dr. Deborah Blacker for many helpful comments used in the preparation of this manuscript.

References

[1] C Anderson, M Nalls, A Biffi, N Rost, S Greenberg, A Singleton, J Meschia, and J Rosand. The effect of survival bias on case-control genetic association studies of highly lethal diseases. *Circulation. Cardiovascular genetics*, 4(2):188–196, 2011.

[2] M Austin, D Simon, and R Betensky. Computationally simple estimation and improved efficiency for special cases of double truncation. *Biometrika*, accepted for publication.

[3] R Betensky, D Louis, and J Cairncross. Analysis of a molecular genetic neuro-oncology study with partially biased selection. *Biostatistics (Oxford, England)*, 4(2):167–178, 2003.

[4] M Bots, A Hoes, P Koudstaal, A Hofman, and D Grobbee. Common carotid intima-media thickness and risk of stroke and myocardial infarction: the rotterdam study. *Circulation*, 96(5):1432–1437, 1997.

[5] CDC. Prevalence of stroke–united states, 2006-2010. *MMWR*, 61(20):379–382, 2012.

[6] V Feigin, CM Lawes, D Bennett, S Barker-Collo, and V Parag. Worldwide stroke incidence and early case fatality reported in 56 population-based studies: a systematic review. *Lancet neurology*, 8(4):355–369, 2009.

[7] D Hallman. The folly of being comforted. *Circulation. Cardiovascular genetics*, 4(2):108–109, 2011.

[8] M Hernan and JM Robins. *Causal Inference*. Chapman and Hall/CRC, to be published.

[9] G Hill. Neyman's bias re-visited. *Journal of Clinical Epidemiology*, 56, 2003.

[10] Y Ino, R Betensky, M Zlatescu, H Sasaki, D Macdonald, A Stemmer-Rachamimov, D Ramsay, J Cairncross, and D Louis. Molecular subtypes of anaplastic oligodendroglioma: implications for patient management at diagnosis. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 7(4):839–845, 2001.

[11] S Johnston, S Mendis, and C Mathers. Global variation in stroke burden and mortality: estimates from monitoring, surveillance, and modelling. *Lancet neurology*, 8(4):345–354, 2009.

[12] E Martin and R Betensky. Testing quasi-independence of failure and truncation times via conditional kendall's tau. *Journal of the American Statistical Association*, 100, 2005.

[13] J Neyman. Statistics; servant of all sciences. *Science (New York, N.Y.)*, 122(3166):401–406, 1955.

[14] S Ostwald, J Wasserman, and S Davis. Medications, comorbidities, and medical complications in stroke survivors: the cares study. *Rehabilitation nursing : the official journal of the Association of Rehabilitation Nurses*, 31(1):10–14, 2006.

[15] KJ Rothman, S Greenland, and TL Lash. *Modern epidemiology*. Lippincott Williams and Wilkins, 2008.

[16] D Sackett. Bias in analytic research. *Journal of chronic diseases*, 32(1-2):51–63, 1979.

[17] WY Tsai. Testing the assumption of independence of truncation time and failure time. *Biometrika*, 77(1):169–177, 1990.

[18] B Turnbull. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Ser. B*, 38:290–295, 1976.

Research participant compensation: a matter of statistical inference as well as ethics

David M. Swanson and Rebecca Betensky

Abstract

The ethics of compensation of research subjects for participation in clinical trials has been debated for years. One ethical issue of concern is variation among subjects in the level of compensation for identical treatments. Surprisingly, the impact of variation on the statistical inferences made from trial results has not been examined. We seek to identify how variation in compensation may influence any existing dependent censoring in clinical trials, thereby also influencing inference about the survival curve, hazard ratio, or other measures of treatment efficacy. We propose a model for how compensation structure may influence the censoring model. Under existing dependent censoring, we estimate survival curves under different compensation structures and observe how they affect the bias of the curve. If the compensation structure affects the censoring model and dependent censoring is present, then variation in that structure affects the accuracy of estimation and inference on treatment efficacy. We illustrate the association between compensation and censoring time under one model. However, as long as compensation affects censoring, and censoring is associated with the event time (i.e., dependent censoring), variation in compensation will result in variation in (biased) inference. From the perspectives of both ethics and statistical inference, standardization and transparency in the compensation of participants in clinical trials is warranted.

1 Introduction: motivation and assumptions

In recent years, there has been increased pressure on investigators to disclose their financial stakes in clinical trials due to concerns over conflicts of interest [9]. Despite this emphasis on transparency, there are no requirements for disclosure regarding payments to research subjects or compensation of investigators for accrual and retention of trial participants. We investigate how variation in these incentives could affect statistical inference through their effect on the retention and drop-out (i.e., censoring) processes. While it is possible that incentives also affect the event time, this is less likely because the event time is often a function of physiological processes. Much of our examination hinges on a background of dependent censoring, i.e., an association between event and censoring times. This is a common feature of clinical trials and impedes valid time-to-event analysis. No attention has been paid in the literature to how incentive structures influence and interact with dependent censoring. However, in the case of some statistical tests, dependent censoring

is not even required for an association between incentives and censoring to invalidate inference [1]. In particular, variation in incentives across trials for research participants, as well as for investigators, may contribute to corresponding variation in participant censoring both dependent and independent patterns and, consequently, inference. In this paper, we use the terms incentive," compensation," and reward" interchangeably, all in reference to the payment used to minimize participant drop-out in clinical trials. It is also important to emphasize at the outset of this investigation that, while we posit that payment may be the chief source of unaccounted-for variation across trials, we can consider payment more generally. For example, it may be a psychological reward such as personal encouragement for retention and compliance from an investigator or caregiver who may be compensated or personal satisfaction from adherence to cultural norms that are specific to study site (this is especially true for international, multi-center trials).

Compensation for subjects varies across trials in its magnitude and in its distribution over the course of the trial and can affect the censoring model in these two ways. Suppose that subjects in a trial who have early event times are inherently more likely to drop out of the trial before their event times are observed than subjects with late event times, so that dependent censoring is present. If there exists an incentive sufficiently generous to convince all subjects to stay on the trial, an unbiased estimate of the survival curve will be available simply because everyone is observed, despite the underlying propensity for worse-performing subjects to be censored. However, if the incentive in a different trial is modest, some proportion of subjects will be informatively censored and the resulting survival curve will be biased. Suppose that subjects in one trial are offered an incentive that is allocated early in the trial in an attempt to shorten the accrual period, which tapers later in the trial to minimize cost, while subjects in a different trial are offered an incentive that is allocated later in the trial to maximize trial retention and completion. If near- term incentives affect subjects' follow-up differently than future incentives, heterogeneity in the censoring distributions across trials is introduced.

In this paper, we first summarize current compensation practices. We then summarize ethical perspectives on the compensation of research subjects. We discuss the potential impact of varying reward structures on cross-trial comparisons and analyses. We posit one plausible censoring model that is a function of the incentive structure and other factors, and through simulation, investigate the variation in estimation of the survival distribution. We demonstrate that reward structures may influence estimated treatment efficacy and so it is important to incorporate this information in statistical analyses and disclose it in the reporting of trials. We encourage journals to add this reporting requirement along with those for conflict of interest and registration of clinical trials.

2 Current clinical trial compensation practices

The few papers that have been published on practices of research subject inducement and compensation in clinical studies reveal variation and a lack of structure in the incentives subjects receive for procedures. Grady et al. [7] surveyed practices of subject payments in 2005 and found that of the 467 surveyed studies, 78% did not specify amount of payment per procedure in the study protocol or subject consent document, and 71% did not specify payment per hour or visit. Some procedures, such as endoscopy, showed little variation in remuneration (consistently $100), while others had considerable variation, such as MRIs ($25-$120) and venipuncture ($10-$50). When payment per visit was recorded, there was variation from $10 to $250. From a broader perspective, compensation for an entire study varied from $5 to $2000, though that variation is partly a function of the study time required and procedure invasiveness. Sixty-six of the studies surveyed were multi-site and 85% of them showed variation in payment across sites. The range of variation was as large as $1000, and the mean and median of inter- site variation were $228 and $120, respectively.

Dickert et al. [4] performed their own analysis of payment practices by investigating 32 research organizations, which spanned the academic, pharmaceutical, contract research organization (CRO), and independent institutional review board (IRB) sectors. The researchers found that only 37.5% of these organizations had specific policies in place regarding payment of research subjects, and as a result, there was no standard of compensation. Indeed, only 18.8% of organizations could even give a confident estimate of the proportion of their studies that were paid. There was also variation in how organizations viewed payment, be it as an incentive to participate (58% of organizations) or compensation for time, inconvenience, or risk. Half of the written guidelines for payment that the investigators reviewed explicitly stipulated that risk should not be compensated. One-fourth of the organizations surveyed had formulas for payment, though the level of specificity varied. When an hourly rate was specified, it varied from $4 to $10. Some organizations paid by day or visit, and payment varied from $25 to $125.

3 Compensation from an ethical perspective

Compensation of research subjects in clinical trials has been an active area of discussion in the ethics literature. Two questions posed in the literature are whether it is ethical for different research subjects to be incentivized differently for the same procedure, and what constitutes "undue" inducement to participate in research, the latter of which is forbidden by the U.S. Common Rule for the Protection of Human Subjects. Here we summarize thinking on both these and other topics. Informed consent is the backbone of ethical conduct of clinical trials, and one aspect of it is the voluntary choice of research subjects to participate [6]. Two elements of voluntariness include the absence of both coercion and "undue inducement" for participation.

Grady [6] argued that physical coercion is not an issue in trials, and current payment methods do not constitute "undue inducement" to participate, but rather should be viewed as compensation for playing a necessary role in a successful clinical trial. Macklin [12] noted that "due" versus "undue" inducement is a relative concept, as what might be due inducement to a lawyer is undue inducement to a minimum wage worker. Thus, identification of an ethical threshold of payment is difficult.

Some authors have approached the definition of undue inducement from the perspective of labor relations and have advocated a wage-payment model of compensation [5]. The tenets of this model are justice, equity, and caution to not make payment so appealing that it is coercive. In contrast, McNeill [13] rejected the idea of research subject payment outright, arguing that the relationship between investigator and subject is unlike that between worker and employer due to the unknown nature of the risks involved. As a result of payment, subjects are unable to adequately assess risks. He argued that since investigators may value their research interests above the well-being of subjects, it is best that subjects are left to volunteer.

While McNeill [13] saw value in individual autonomy, he argued that equitable safeguards come prior to those rights. Even though Macklin [12] recognized the danger of paternalism, she, like McNeill, saw a role for providing safeguards that are in the subjects own best interests. She suggested that inducements should err on the low side and that a study may be deemed unethical if an inadequate number of research subjects are not inclined to participate in those circumstances. To that same end, Grady [6] argued that payment should be standardized so that it is similar to that of other unskilled labor in the surrounding community.

Lemmens and Elliot [10] argued that the nature of the ethical relationship between a research subject and an investigating institution depends on the health of the subject. Subjects who are ill and therefore stand a chance of receiving therapeutic benefit from experimental treatments are in a vastly different position than healthy subjects whose primary motivation for participation is compensation. In the latter case, Lemmens and Elliot, along with other writers, maintained that the relationship should be viewed as that of a labor contract and have similar protections.

4 Impact of varying reward structures

If payment or some other factor altered the underlying dependent censoring mechanism, one would expect to see variation in outcomes of the placebo arms of trials of similar patient populations if incentive variation is present. Schneider and Sano [17] summarized cognitive decline as measured by ADAS-Cog in placebo arms of phase II and III Alzheimers drug clinical trials. They reported that even for moderate sample sizes of $107 - 317$, the range of the mean decline over 18 months was 4.3 to 9.1 points. Although it is not possible to definitively identify the cause for the placebo group variation, differences in participant demographics

across Alzheimers drug trials does not seem to explain the observed variation since study subjects had similar ages, education levels, APOE e4 genotypes, baseline ADAS-Cog scores, and there was consistency in study eligibility criteria [17]. It would be useful to know what incentives were provided to caregivers and site investigators associated with these trials. Interestingly, two different trials of the same drug and similar design had mean cognitive declines closer to one another, suggesting that, be it through consistency of payment or other factors, trials with common features may cause similar subject behavior.

When unexplained variation in the placebo arms of trials is present, it raises doubts about the validity of meta-analysis. In the case of Alzheimers disease, if the variation in placebo arm mean cognitive decline were reflective of biased estimation, the results would not be strengthened by meta-analysis since combining biased results does not diminish bias. With increasing pressure for open access to information and data for public dissemination and meta-analysis, this issue should be recognized. For example, the Alzheimers Disease Cooperative Study (ADCS) provides data from placebo arms of Alzheimers clinical trials for public use (www.adcs.org), the Prize4Life foundations Amyotrophic Lateral Sclerosis (ALS) Pro-Act database provides data from placebo arms of ALS clinical trials (www.prize4life.org), and the CDCs Tuberculosis Trials Consortium (TBTC) gives investigators involved in the consortium access to the data (www.cdc.gov/tb/topic/research/tbtc). If the incentive structures used in these studies impacted the retention to the studies, and if they are not known and adjusted for in the meta-analyses, any conclusions are likely to be invalid.

5 Simulation study

We illustrate through simulation how variation in research participant compensation could lead to variation in inference on the survival curve. We consider a particular model for censoring in order to encode the relationship between censoring and compensation and to investigate how that relationship might further influence the underlying dependent censoring. We estimate the survival curve using the Kaplan-Meier estimator [8] under various compensation structures and graph the results to demonstrate

the variability of the curve (see Figure 15). We do not graph the true event survival distribution in any of the figures in order to emphasize the variability in estimation introduced by different incentive structures and because bias is already present when there is dependent censoring, regardless of the incentive structure.

6 Event and censoring models

We simulated clinical trials with $40,000$ subjects. Half of the subjects had covariate Z_i equal to 0 and half had Z_i equal to 1. For example, Z_i might be socio-economic status (SES), which affects both subjects event

times (since SES is associated with health) and censoring times (since those in higher SES strata may stay on the trial longer or shorter, depending on the situation). In simulations used to generate Figure 15, the event times of those with $Z_i = 0$ were distributed as Gamma$(6, 7/12)$ (shape parameter 4 and rate parameter 5/6), and those with $Z_i = 1$ were distributed as Gamma$(4, 5/6)$. In all simulations, we rounded up the generated event times to make them discrete. The discrete event times are interpreted as the time points at which a subject was observed to have had an event.

A subjects event time was observed if it occurred prior to the minimum of the subjects randomly generated censoring time and 10 time units. The model for the discrete time hazard of censoring at time j for subject i for all simulations was

$$\text{logit } P(\text{Subject}_i \text{ censored at time } j | \text{Not censored prior to time } j, Z_i) = \alpha_j + \mu \cdot m_j + \gamma \cdot Z_i + \rho \cdot j$$

where $\text{logit}(p) = \log(p/(1-p))$, α_j helps control the overall proportion of censoring at each time period j, γ is a constant across time periods and subjects and controls to what extent the covariate Z_i influences the censoring hazard, ρ controls the degree to which subjects become "tired" of being involved in the trial and want to drop-out, μ controls the influence of the incentive m_j on the hazard of censoring, and m_j is a function of the incentive structure over the remaining time periods in the trial and the "incentive anticipation" parameter (or equivalently, decay parameter) described below. We define m_j as $m_j = \sum_{k=j}^{tot} incentive_k \cdot dec^{(tot-k)}$, where tot is the total number of follow-up periods (10 in our case), $incentive_k$ is the incentive size at time period k, and dec is the decay or compounding parameter (0.88 in our case). Dependent censoring is present in the simulation because covariate Z_i affects the hazard-of-censoring (through γ) and the event time (because the gamma distribution parameters depend on Z_i), and we do not condition on the covariate when estimating the survival curve.

The total amount of the incentive for all of the structures shown is 28 units, but those units are distributed in three different ways for the three survival curves shown in Figure 15. Each survival curve corresponds to a certain incentive structure, and that corresponding incentive structure is shown in Figure 16. Under all structures, payout of the 28 units is static at a level of two in 8 of the 10 follow-up periods, but then jumps to a higher payout of six for two consecutive follow-up periods at different times during the simulated trial. A structure with an early jump represents a trial whose investigator wants quicker subject accrual, while the structure with a later jump represents a trial whose investigator desires complete follow-up. Parameter values for all simulations were $\alpha = (18.5, 16.5, 14.5, 11.5, 11.5, 11.5, 8.5, 8.5, 8.5, 8.5, 8.5)$, $\mu = 1$, $\gamma = -8/3$, and $\rho = 1/60$.

7 Results

Figure 15 depicts downward biased survival curves calculated with the Kaplan-Meier estimator [8] under the different incentive structures shown in Figure 16, where there is a color correspondence between the estimated survival curve and the incentive structure used in the associated censoring model. Differences in the survival curves in Figure 15 illustrate that when dependent censoring is present and the incentive structure affects the censoring model, variation in incentive structure results in variation in the curve.

While the simulation is conducted under a particular censoring model, the qualitative role of the incentive within the model is highly plausible. Indeed, an association between incentive structure and censoring time is the only assumption needed for this discussion to be relevant; in the presence of dependent censoring, incentives are a source of significant variation that could lead to variation in inference via the censoring model.

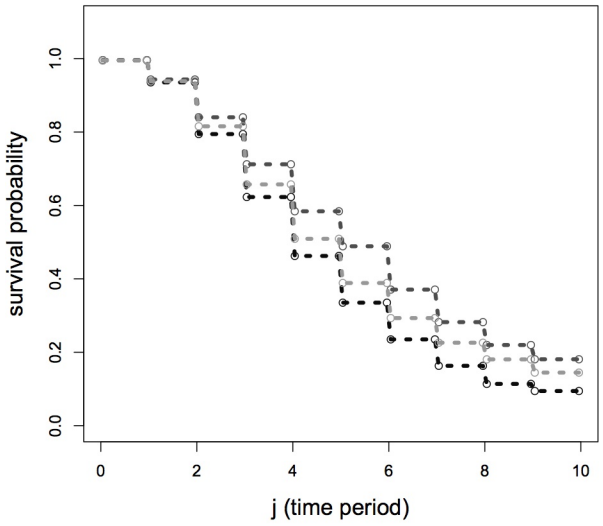

Figure 15: Survival curves generated under the different incentive structures shown in Figure 16.

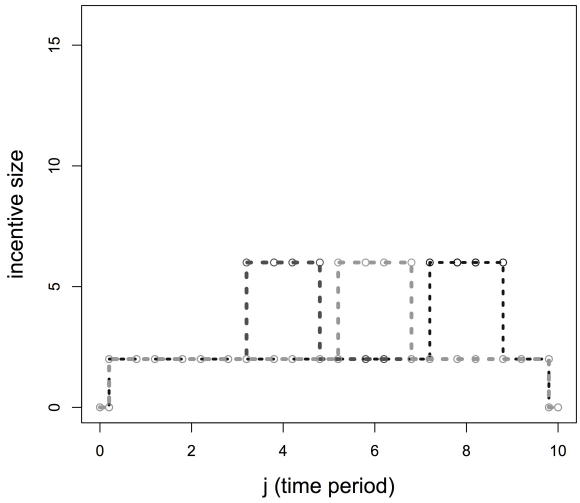

Figure 16: Incentive structures corresponding to the survival curves shown in Figure 15.

8 Adjustment for incentives

Having established the impact of heterogeneity in incentive structure across trials, or across sites within a single trial, we turn to an examination of potential analytic adjustments that might be made. We first note that adjustment for trial or site in an analysis would not be sufficient to account for inter-trial or inter-site variation since the issue invalidating inference is neither a main effect of trial or site, nor clustering

of observations, but rather potential violation of the event and censoring time independence assumption, fundamental to most time-to-event analyses. Sensitivity analysis is a good way to assess how analysis assumptions influence inference. An Institute of Medicine study [2] was commissioned to study the problem of missing data in clinical trials, and two associated articles have highlighted the importance of sensitivity analysis in the context of missing data, of which censored survival data is a subcategory [11, 20]. The Institute of Medicine report [2] in particular recommended specific ways of conducting sensitivity analysis with time-to-event data. One such way is to use auxiliary prognostic factors to explain residual dependence between event and censoring times within strata of model covariates. If these prognostic factors are not sufficient to explain the dependence, other frameworks are suggested in which a non- identifiable censoring bias parameter that encodes residual dependence can be manipulated to investigate sensitivity of inference to it [16, 2]. Regardless of the context in which sensitivity analysis is performed, both the Institute of Medicine report and corresponding New England Journal of Medicine editorial state that it is important to know how robust analysis findings are to missing data assumptions [2, 20]. Shih [18] also recognized the importance of sensitivity analysis in clinical trials in which dependent censoring may be present, though recommended assessing sensitivity using a broader class of missing data approaches, such as multiple imputation and non-parametric rank-based methods. Since it is only when the assumption of independent censoring and event times is violated that varying payment structures may lead to variation in survival curve or hazard ratio estimates, correcting for its possible violation is important. Correction for dependent censoring and analysis of sensitivity to the assumption of independent censoring can be done using a weighting scheme described in Robins (1993) and Robins and Finkelstein (2000). The weight placed on any individual at time t is the inverse of the probability of being censored by time t. Thus, those who are more likely to be censored are weighted more in a given model, the idea being that they are "stand-ins" for subjects similar to themselves who were unobserved because of a high hazard of being censored. Weights are calculated by modeling the hazard of censoring within strata of the covariates already included in the hazard or survival model of interest using additional covariates that could possibly predict censoring. These weights can then be used in the inverse probability weighted (IPW) analogue of the Kaplan-Meier estimator:

$$S_T(t|z) = \prod_{\{i;X_i<t\}} 1 - \frac{\tau_i W_i(X_i) I(Z_i = z)}{\sum_{k=1}^{n} Y_k(X_i) W_k(X_i) I(Z_k = z)}$$

where τ_i is the event indicator for subject i, Z_i is the covariate or indicator of stratum, $Y_i(t)$ is an indicator for presence in the risk set at time t, $I(\cdot)$ is the indicator function, and X_i is the observed event or censoring time. The weight, $W_i(t)$, is set to 1 in the case of the typical Kaplan-Meier estimator, and is the inverse of the probability of remaining uncensored up to time t for the weighted Kaplan-Meier estimator. That probability

can be estimated in different ways, including by fitting a Cox model [3] within strata of Z_i, conditional on covariates that may influence censoring. In our case, weights were obtained by first calculating the proportion of censored individuals within each stratum of the covariate at each time period j among the risk set, giving the hazard of censoring, $h_j(Z_i)$. With this information, probability of remaining uncensored to time t given the covariate was calculated by taking the product of one minus the hazards up to t: $P_i(t) = \prod_{\{j<t\}}(1 - h_j(Z_i))$. Weights were calculated using $W_i(t) = 1/P_i(t)$ so that individuals more likely to be censored by time t were weighted more than those less likely to be censored. We see in Figure 17 that when dependent censoring is present and there are two different incentive structures, there are two corresponding and different survival curves estimated, though the underlying event model is identical for both curves. However, after correcting for the dependent censoring using weights, we observe in Figure 18 that there is far less variation between the two estimated survival curves because the estimator is unbiased and therefore its expectation is unaffected by incentives. One might propose to somehow adjust for varying incentives themselves to reduce estimate variation, though proper adjustment would require knowledge of the relationship between incentives, other covariates, and the censoring hazard. Apart from such knowledge, there is no clear method for adjustment and so we focus our correction recommendation on the underlying issue of dependent censoring. If there are multiple trials of an identical therapy in the same participant population with dissimilar survival curves for the placebo arms, it might indicate that dependent censoring is present and different payment structures are affecting the censoring model. We therefore recommend that one should attempt to first calculate weights for each trial by modeling the hazard of censoring using prognostic factors likely to predict it. Then, using the procedure described in Robins and Finkelstein (2000), one should calculate a weighted Kaplan-Meier survival curve for each trial. If the placebo arms of the trials show less variability across trials, which should be the case if entry criteria are similar and sample sizes are adequate, correction for dependent censoring may be sufficient and treatment efficacy within each trial can be analyzed. While there is no way to test for the presence of dependent censoring in either the weighted or unweighted Kaplan-Meier estimator, such a procedure may give researchers additional insight into clinical trials data and the robustness of their results to modeling assumptions. If one can correct for the dependent censoring that is present, then in theory survival curves estimated across trials under different payment structures should be equivalent since they estimate the same, underlying, event model. As a result, assuming that the censoring distributions differ across trials, even if dependent on event, one has the ability to check for whether the correction for dependent censoring is done well. This is an advantage afforded by the multiplicity of trials or sites within a trial, as dependence of censoring and event is generally not testable [19]. This result is a general one: if dependent censoring is present and survival curves are estimated under different censoring models (all of which are associated with the event model) so that the estimated survival curves are

all different, then correcting for the dependent censoring will give unbiased estimates of the survival curves so that much less variation in those estimates should be observed. If we do not observe a decrease in the variation of the survival curves across trials after weighting, we would know that weights were not calculated effectively. Our emphasis on payment structure and its possible effect on inference would be unnecessary if either it or dependent censoring were accounted for in analyses. However, a review of a variety of terms used to describe dependent censoring in six major medical journals using PubMed and Google Scholar reveals that it is often not discussed. Using these tools to search all available archives of the New England Journal of Medicine (NEJM), Archives of Internal Medicine, Journal of the American Medical Association (JAMA), Annals of Internal Medicine, Nature Medicine, and the Lancet with terms such as informative censoring, dependent censoring, non-informative dropout, and permutations of them returned 65 total articles, with NEJM, JAMA, and the Lancet each contributing approximately 15 articles to the total number. A review of the articles showed that in some cases investigators did consider the possibility of informative censoring being present in their studies and performed analyses under different modeling assumptions. In other cases, investigators performed analyses that might have been less prone to violated assumptions. However, even if all articles found appropriately addressed the possibility of informative censoring, it would still be a fraction of all time-to-event analyses described in clinical journals.

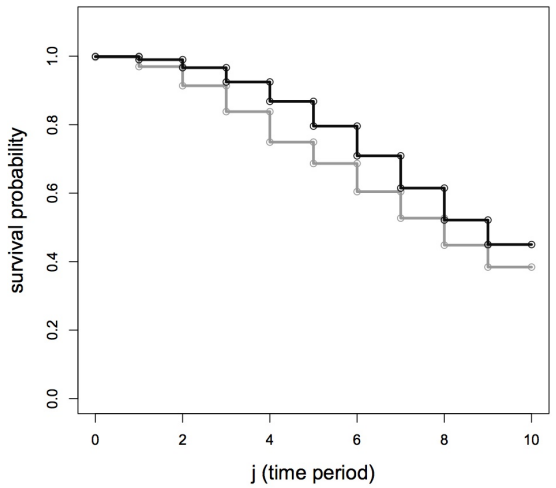

Figure 17: Survival curves calculated under two different incentive structures. There is significant variability in the curves estimated under the different structures.

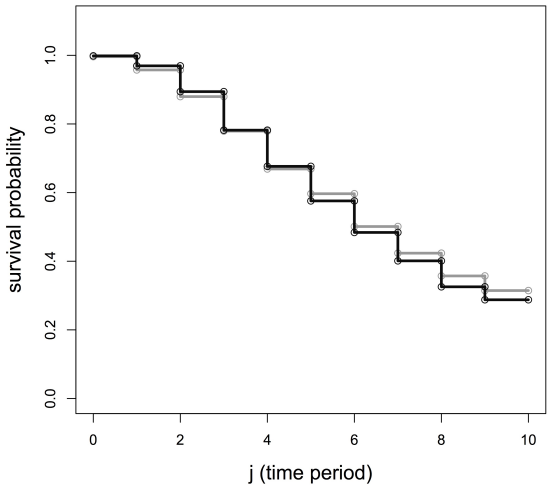

Figure 18: An inverse probability weighted analogue of the Kaplan-Meier estimator gives less variable survival curves under two different incentive structures.

9 Conclusion

We have demonstrated that variation in compensation of research subjects may not only be unethical as claimed in prior literature, but may also have substantial influence on statistical inference in clinical trials. Through its influence on the censoring model of research subjects, different compensation patterns may result in different censoring models and thus change inference if some degree of dependent censoring is present in the trial. We have focused the context of this paper on clinical trials, but variation in compensation can affect observation studies as well. In that setting, different compensation structures would more influence which populations are attracted to participate, though this point is relevant to clinical trials, too, as participant demographics may vary with compensation, both at baseline and across time points in the trial. While some might simply advocate for more compensation of research subjects so that there is a high percentage of complete follow-up among participants, this suggestion is not feasible from both ethical and financial standpoints. For consistency across clinical trials, standardization in compensation is essential. While standardization is difficult to achieve both because incentives effectiveness at increasing participant retention is a function of participant values, demographics, socioeconomic status, and psychology, and not all incentives are monetary, some greater degree of standardization is likely helpful to decrease censoring distribution variability. However, even with standardized compensation, estimates of treatment efficacy are likely to be biased in the presence of dependent censoring. Thus, use of statistical methods that correct for dependent censoring and adjust for variable compensation structure when it is present should be emphasized more in the clinical literature. We encourage clinical journals to mandate the application of these methods, along with full disclosure of the compensation of subjects, caregivers, and investigators.

10 Acknowledgments

The authors wish to thank Dr. Deborah Blacker in the preparation of this manuscript, who offered many helpful comments and ideas.

References

[1] Norman Breslow. A generalized Kruskal-Wallis test for comparing k samples subject to unequal patterns of censorship. *Biometrika*, 57, 1970.

[2] National Research Council. *The Prevention and Treatment of Missing Data in Clinical Trials*. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press, panel on handling missing data in clinical trials edition, 2010.

[3] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972.

[4] N. Dickert, E. Emanuel, and C. Grady. Paying research subjects: an analysis of current policies. *Annals of Internal Medicine*, 136(5):368-373, 2002.

[5] N. Dickert and C. Grady. What's the price of a research subject? approaches to payment for research participation. *New England journal of medicine*, 341(3):198-203, 1999.

[6] C. Grady. Money for research participation: does it jeopardize informed consent? *American journal of bioethics*, 1(2):40-44, 2001.

[7] C. Grady, N. Dickert, T. Jawetz, G. Gensler, and E. Emanuel. An analysis of US practices of paying research participants. *Contemporary clinical trials*, 26(3):365-375, 2005.

[8] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[9] Aaron Kesselheim, Christopher Robertson, Jessica Myers, Susannah Rose, Victoria Gillet, Kathryn Ross, Robert Glynn, Steven Joffe, and Jerry Avorn. A randomized study of how physicians interpret research funding disclosures. *The New England journal of medicine*, 367(12):1119–1127, 2012.

[10] T. Lemmens and C. Elliott. Guinea pigs on the payroll: The ethics of paying research subjects. *Accountability in Research*, 7(1):3-20, 1999.

[11] Roderick Little, Ralph D'Agostino, Michael Cohen, Kay Dickersin, Scott Emerson, John Farrar, Constantine Frangakis, Joseph Hogan, Geert Molenberghs, Susan Murphy, James Neaton, Andrea Rotnitzky, Daniel Scharfstein, Weichung Shih, Jay Siegel, and Hal Stern. The prevention and treatment of missing data in clinical trials. *The New England journal of medicine*, 367(14):1355–1360, 2012.

[12] R. Macklin. 'Due' and 'Undue' Inducements: on passing money to research subjects. *IRB: Ethics and Human Research*, 3(5):1-6, 1981.

[13] P. McNeill. Paying people to participate in research: why not? *Bioethics*, 11(5):390-396, 1997.

[14] J Robins and D Finkelstein. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics*, 56(3):779–788, 2000.

[15] JM Robins. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. *Proceedings of the Biopharmaceutical section, American Statistical Association*, pages 24–33, 1993.

[16] Daniel O Scharfstein and James M Robins. Estimation of the failure time distribution in the presence of informative censoring. *Biometrika*, 89(3):617–634, 2002.

[17] Lon S Schneider and Mary Sano. Current alzheimer's disease clinical trials: methods and placebo outcomes. *Alzheimer's & dementia*, 5(5):388–397, 2009.

[18] Weichung Shih. Problems in dealing with missing data and informative censoring in clinical trials. *Current controlled trials in cardiovascular medicine*, 3(1), 2002.

[19] Anastasios Tsiatis. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72(1):20–22, 1975.

[20] James Ware, David Harrington, David Hunter, and D'Agostino, Ralph. Missing data. *The New England journal of medicine*, 367(14):1353–1354, 2012.